Analysis and treatment of the Søndersø time series

Grey Box Well Field Modelling

Dorini, Gianluca F.; Thordarson, Fannar Ørn; Madsen, Henrik; Madsen, Henrik

Publication date:
2011

Document Version
Publisher's PDF, also known as Version of record

Grey Box Well Field Modelling

Analysis and treatment of the Søndersø time series

Gianluca F. Dorini[1]
Fannar Örn Thordarson[1]
Henrik Madsen[1]
Henrik Madsen[2]

Technical Report No. 4

March 31, 2011

Informatics and Mathematical Modelling
Technical University of Denmark

[1] DTU Informatics; Technical University of Denmark
[2] DHI, Water • Environment • Health

Contents

Introduction and outline

This report deals with grey box modelling applied to the Well Field Optimisation project. The subject is the real case study of Søndersø, located north-west of Copenhagen, DK. This report contains a comprehensive description on how the dataset of measurements taken at Søndersø have been treated and analysed. The purpose of such analysis is twofold. Firstly is to identify a suitable architecture for the grey-box model. Secondly to design a procedure to select values from the dataset that will be used for the calibration of the parameters of the grey-box model. Section 1 describes the Søndersø well field, and provides an overview of the dataset. Section 2 describes the numeric treatments that have been applied to the dataset; the result is summarized in Section 3. Section 4 illustrates the analysis performed on the treated dataset. In this section, the fundamental mechanisms of the well field system are detected and decomposed (subsections 4.1 - 4.3). Based on the results of such analysis, a simple modelling exercise is performed showing that linear models can be effectively employed to simulate a well field (subsection 4.4). Section 5 describes a sampling approach, designed to calibrate the parameters of the grey-box model with a representative database which is also reasonably reduced in size. Summary and conclusions are in Section 6.

1 Dataset overview

The dataset provided contains three time series of measurements taken at Søndersø well field. The Søndersø well field, displayed in Figure 1, consists of 21 wells divided into three groups; West (3), East (8), and South (10). Wells are all interconnected through a water distribution network (WDN), each aforementioned group is a branch of the WDN. Each time series is a sequence of measures taken at regular interval for several variables.

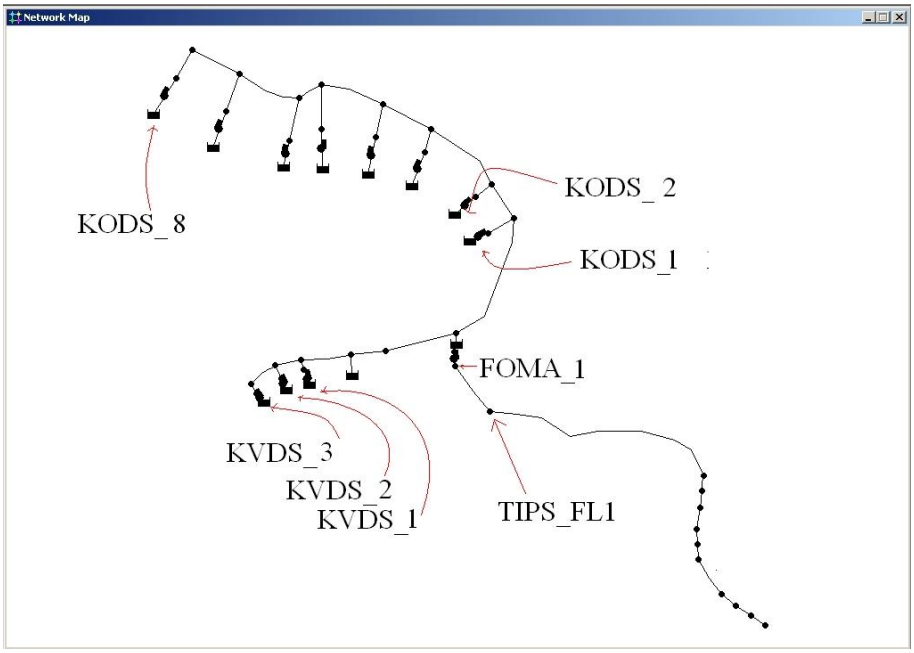


Figure 1: The Søndersø Well Field

The variables are:

- The pumping rates at well level for the East and West branch, $q = (q_t^1, q_t^2, \ldots, q_t^{11})$. In the provided files, pumping rates of the east branch of wells, $q_t^1, \ldots, q_t^8$, are labelled as KODS_FL1,..., KODS_FL8. For the west side, $q_t^9, q_t^{10}, q_t^{11}$, the labels are KVDS_FL1 , KVDS_FL2 , KVDS_FL3.

- The total pumping rate of the South branch, q_t^s; labelled as TIPS_FL1

- The total pumping rate of the whole well field, Q_t; labelled as FOMA_1.

- The water drawdown at well level for the East and West branch, $h = (h_t^1, h_t^2, \ldots, h_t^{11})$. In the provided files, water drawdown of the east branch of wells, $h_t^1, \ldots, h_t^8$, are labelled as KODS_DP1,..., KODS_DP8. For the west side, $h_t^9, h_t^{10}, h_t^{11}$, the labels are KVDS_DP1 , KVDS_DP2 , KVDS_DP3.

The three time series cover a period of respectively 140, 80 and 35 days, see Figure 2. The measurements are instant values that have been taken every minute, for a total of 397,625

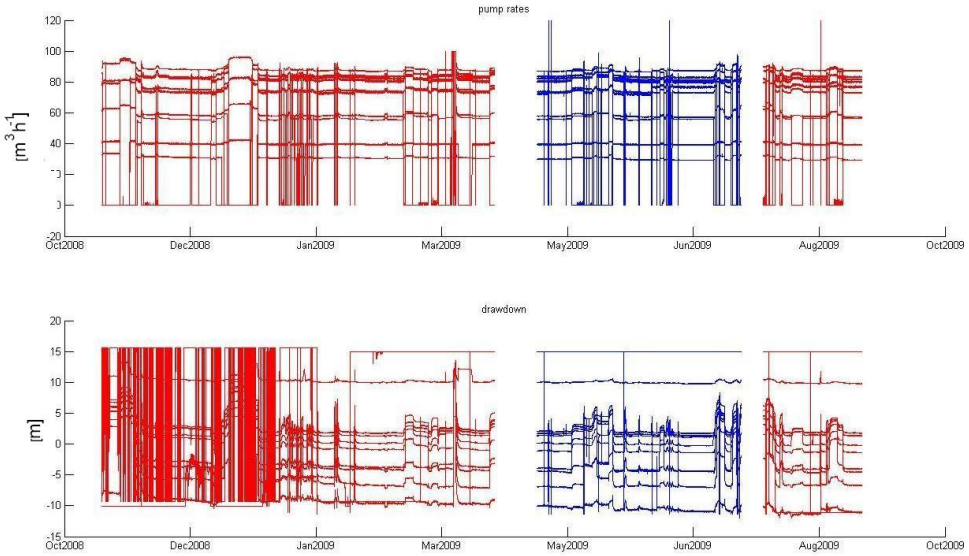


Figure 2: The three provided time series of Pumping rate and water drawdown.

records. The same series have also been provided with one-hour interval. However, such a dataset is nothing but a subset of the one-minute interval dataset, therefore it contains only redundant information.

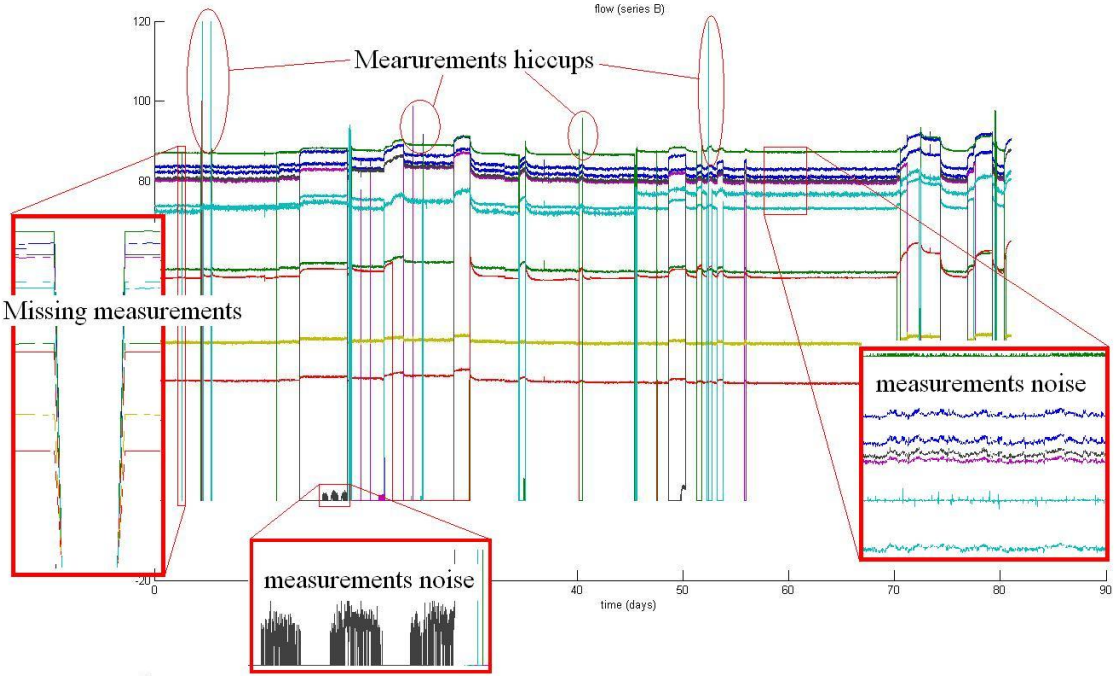
2 Data treatment

Comments of data quality The overall quality of the data provided is satisfactory as most of the records are adequate for modelling purposes. Furthermore, the size of the dataset is large enough to guarantee convincing validation tests. Despite this, the whole dataset contains some flaws, requiring major or minor treatments. This section describes the various types of flaw observed within the data, see Figure 3a and 3b.

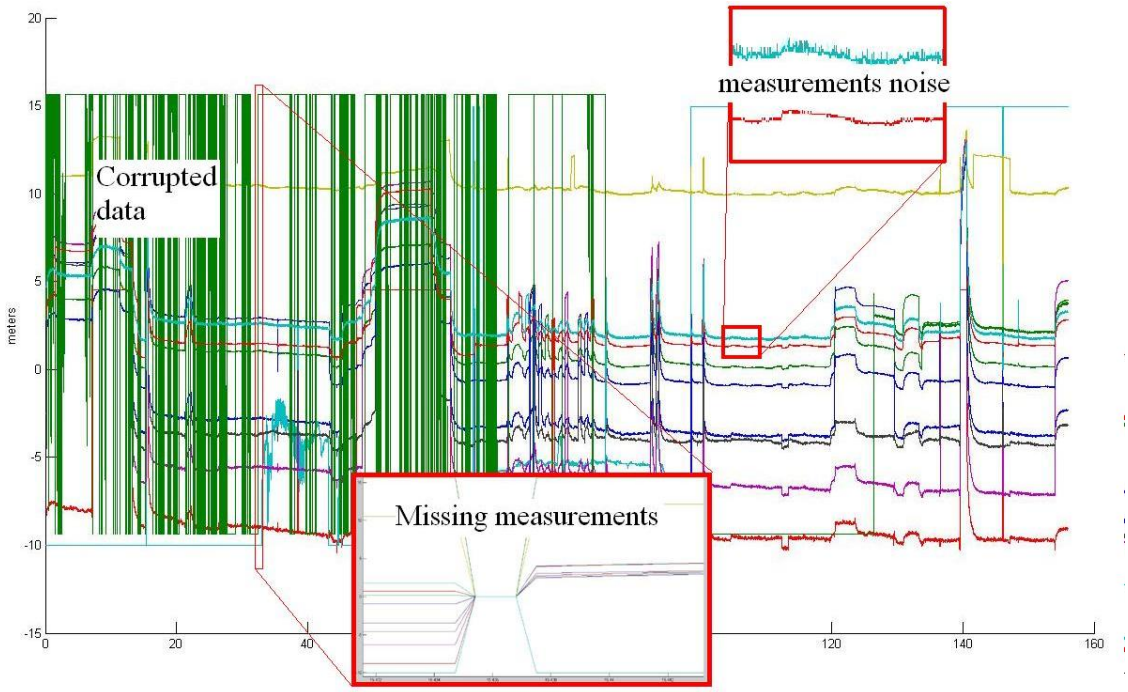
Flow Balance mismatches. Even though each record t in every pumping rate time series should fulfil the mass balance

$$Q_t = q_t^s + q_t^1 + \ldots + q_t^{11}$$

this seems not to be the case in the data provided. The plots of Figure 4 show a discrepancy between Q_t and $q_t^s + q_t^1 + \ldots + q_t^{11}$ in all three time series. The mean difference is around 4%, and it seems to be linearly correlated, as shown in Figure 5. The q^s series has been ignored as it contains redundant information.



(a) *Data flaws in the measured pumping rates.*



(b) *Data flaws in the measured drawdown.*

Figure 3: Overview of data flaws.

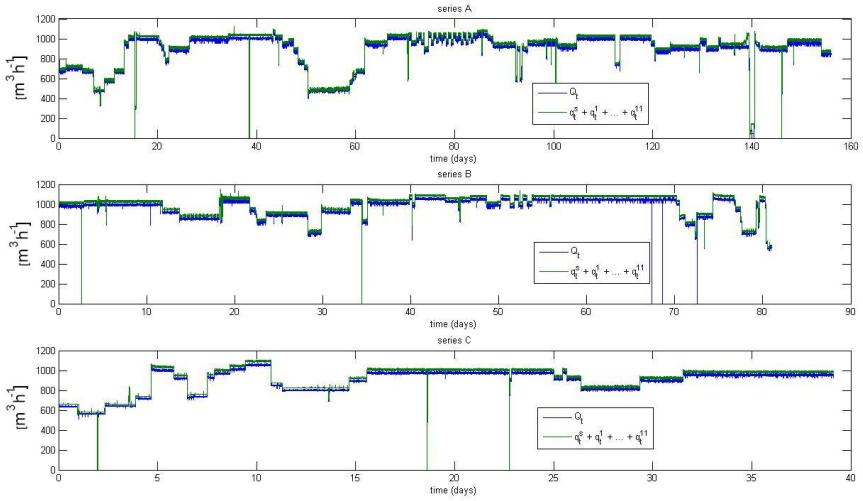


Figure 4: Observed discrepancies in pumping rates.

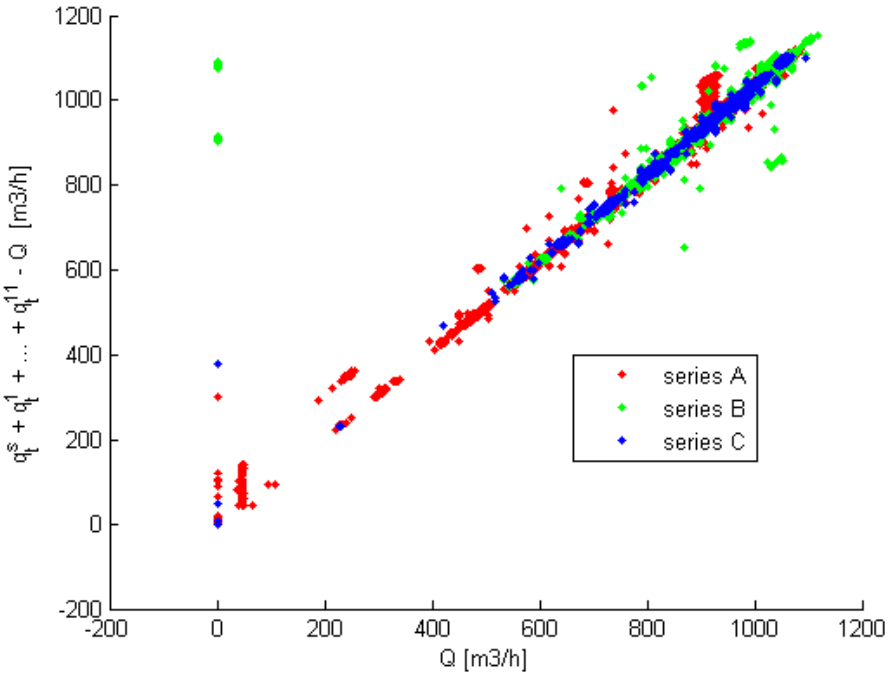


Figure 5: Linear correlation of the discrepancies in pumping rates.

Measurement noises. All time series follow rather smooth time patterns with *noisy* oscillation. In both pumping rates and water drawdown series, the amplitude of such noisy oscillations are on average within few percentage points of the measurements. The noise is certainly caused by the inherent imprecision of the instruments utilised to take the measurements. Such an error will be treated as part of the model. The only

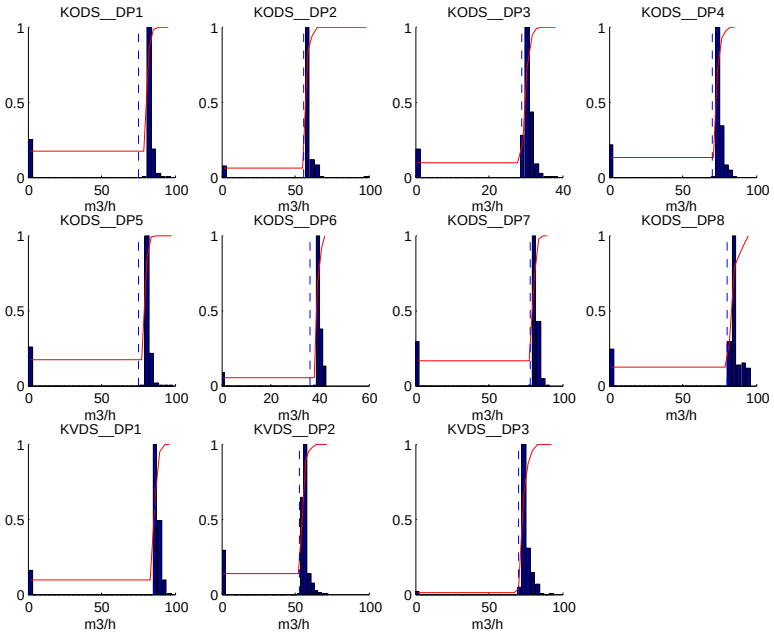


Figure 6: Sampling frequency of pumping rates for the 11 pumps of Søndersø well field.

exception concerns the observed small positive noises in pumping rate measurements. Although pumping rates are not stationary, they vary within relatively narrow ranges. Those ranges differ from pump to pump, as shown in Figure 6. Small positive noises in pumping rate measurements certainly refer to switched off pumps, therefore the values have been set to zero. Furthermore, *small positive noises can be utilised to study the statistical characteristic of the pumping rate measurement errors*.

Missing measurements and measurement *hiccups*. For both pumping rates and drawdown, a case of missing data is when values turn suddenly to zero for all wells and remain zeros only for one or few minutes. Measurements *hiccups* consist in sudden and isolated picks in pumping rates. They are easy to identify because they significantly exceed the normal ranges of the pumping rate values. After a case of hiccups or missing data, the series continue following the same pattern they were following before the event. *The action taken was to manually identify time windows when missing measurements and measurement hiccups occurred, and interpolate the data patterns through such time windows*. See Figure 7.

Corrupted data. Corrupted data are long sequences where measurements are altered and unusable. Here, a distinction is done between the two cases when data corruption affects pumping rate time series and water drawdown time series. The distinction is based on the fact that pumping rates are decision variables, whereas water drawdown levels are state variables. The difference between decision variables and state variables

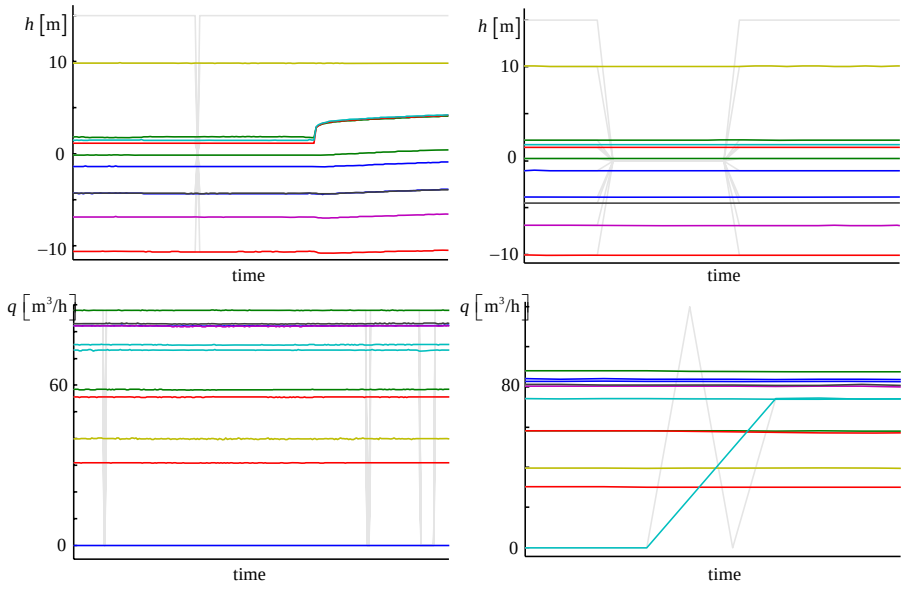


Figure 7: Data interpolation.

is that the latter are controlled by the former and not vice versa. This can be better understood by looking at the partial differential equations governing the groundwater flow. The groundwater flow in three dimensions is expressed as:

$$\frac{\partial}{\partial x}\left(K_x \frac{\partial h}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y \frac{\partial h}{\partial y}\right) + \frac{\partial}{\partial z}\left(K_z \frac{\partial h}{\partial z}\right) + q_v = S_s \frac{\partial h}{\partial t} \tag{1}$$

where K terms are the hydraulic conductivity in each coordinate direction, source-sink term q_v represents flow rate per unit volume, and S_s is the specific storage. In most well field applications these equations are imposed on a finite domain with given boundary conditions. The groundwater flow equation is often in two spatial dimensions,

$$\frac{\partial}{\partial x}\left(T_x \frac{\partial h}{\partial x}\right) + \frac{\partial}{\partial y}\left(T_y \frac{\partial h}{\partial y}\right) + q = S \frac{\partial h}{\partial t} \tag{2}$$

where the T terms are the transmissivity in each coordinate direction, the vertically averaged source-sink term q is the pumping rate, and S is the storage coefficient. If the aquifer is unconfined, then the two-dimensional form equation (2) becomes

$$\frac{\partial}{\partial x}\left(K_x (h-\sigma) \frac{\partial h}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y (h-\sigma) \frac{\partial h}{\partial y}\right) + q = \eta_d \frac{\partial h}{\partial t} \tag{3}$$

where σ is the elevation of the base of the aquifer so that $(h-\sigma)$ gives the aquifer thickness and η_d is the drainage porosity; this unconfined equation is nonlinear in the state variable h. Equations (2) and (3) clearly show that water drawdown h is a state variable, whose dynamic depends on the pumping rate q; the decision variable. Decision variable dynamic is unknown, i.e., decision variables are arbitrary variables. Suppose that between time t_1 and t_2, the water drawdown is corrupted for a well x only, and still intact

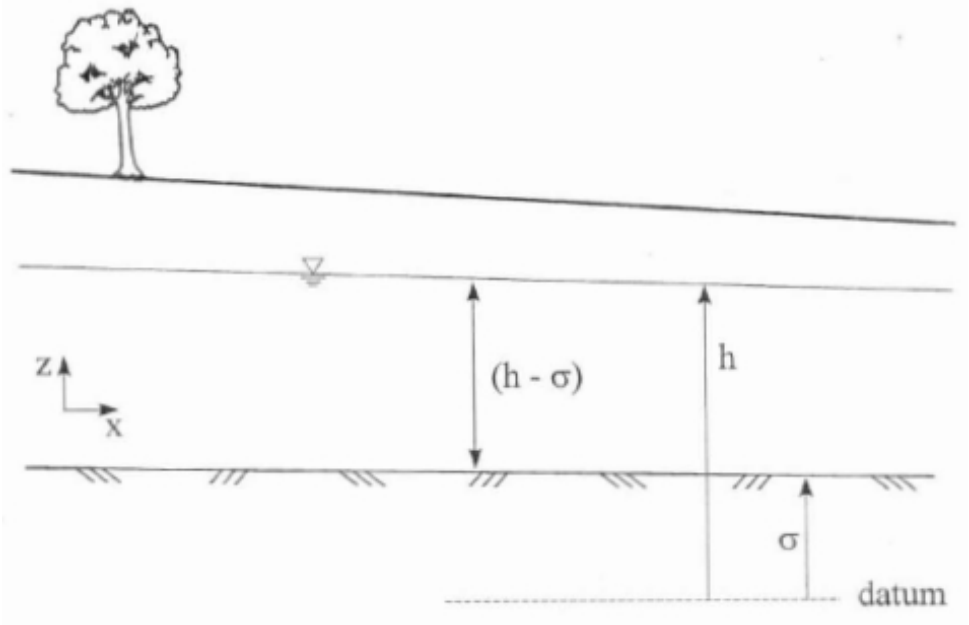


Figure 8: Variables used in the unconfined aquifer equations (Ahlfeld & Mulligan, 2000).

for the remaining wells. Also suppose that during the same period the pumping rate time series are intact for all wells, including well x. In this case the series can still be utilised, as the 10 depending state variables depend on the known decision variables. Clearly, state variable also depend on each other, and the missing information of the *corrupted* well is a limitation. However such a limitation is not striking if the time series between time t_1 and t_2 will be utilised for model validation and not for model calibration. In that case the validated model will be actually used to reconstruct the missing water drawdown level of well x. *The action taken was to manually examine the water drawdown series well by well, and replace them with* `NaN` *values (Not a Number)*. See Figure 9.

Completely different is the opposite situation. Suppose that between time t_1 and t_2, the pumping rate is corrupted for well x only, and still intact for the remaining wells. Also suppose that during the same period the water drawdown time series are intact for all wells, including well x. In this case the time series between time t_1 and t_2, cannot be used even for validation. In fact, the model will still require the pumping rate of well x as input. The model will not be designed to reconstruct missing input variables. *The action taken was to manually detect time intervals where the pumping rate are corrupted for* at least *one well. Those time intervals have been completely removed from the time series.* See Figure 10.

3 Treated data set

As result of the application of the aforementioned methods to the three time series provided, five treated time series have been obtained, see Figure 11. The series are 5 as the

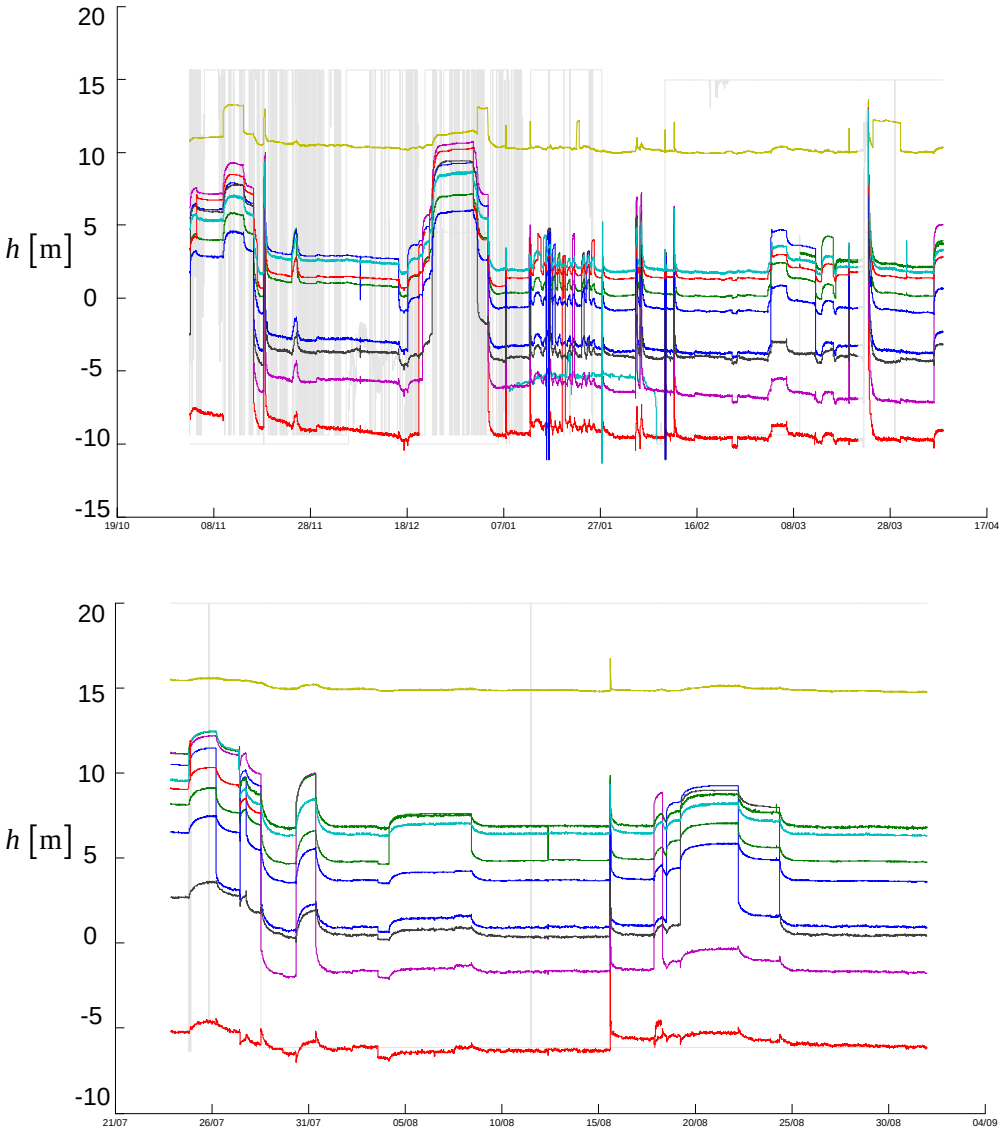


Figure 9: Data elimination

first and second series have been split, due to corrupted pumping rate data. In terms of records, the length of the treated series are respectively, 199,211, 22,293, 114,416, 2,007, and 56,340. The total number of records of the treated series is equal to 394,267; namely 0.84% shorter than the original. Among the selected records, a total number of 609,842 water drawdown measures have been removed (set `NaN`) from a total of 4,336,937 measurements; namely the 14.06%.

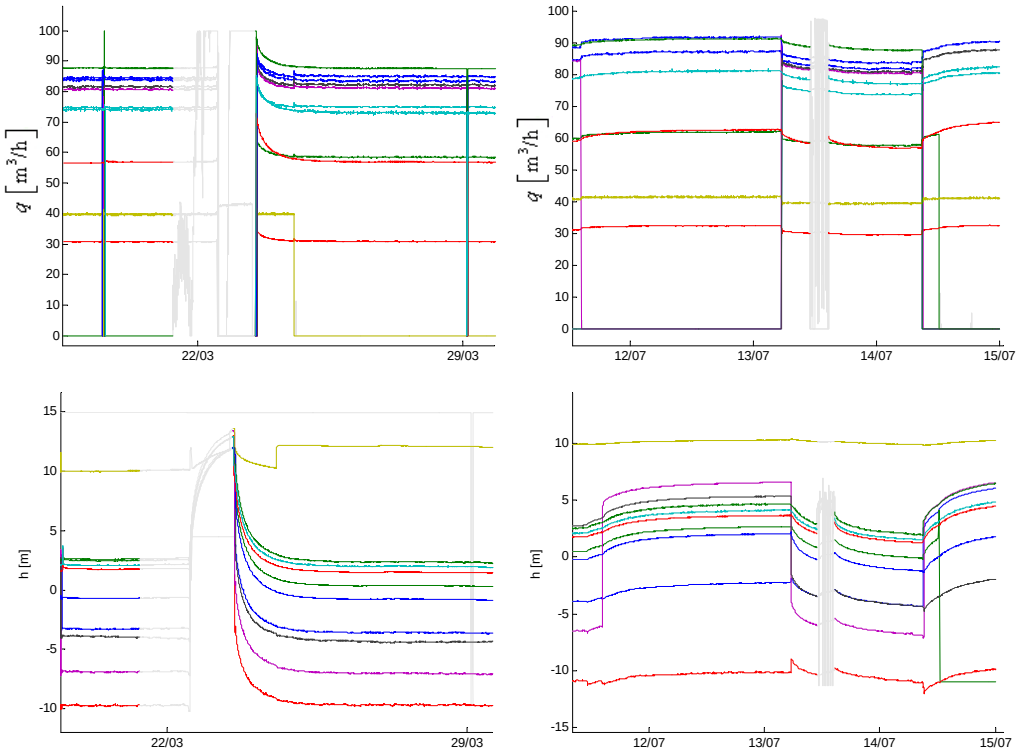


Figure 10: Series splitting

4 Preliminary analysis of the model

4.1 Data partitioning

The various treatments discussed in Section 2 have been applied, the resulting dataset consists on 5 time series (Section 3). At this point, the *cleaned* data are observed, the relationship between decision variables (pumping rates) and state variables (water drawdown) are analysed in a qualitative way. Figure 11 shows what pointed out in section 2 *"Although pumping rates are not stationary, they vary within relatively narrow ranges"*. The pumps installed in Søndersø well field are all of the ON-OFF type, except for the three pumps installed in the west part, i.e. 9,10,11. However, data do indicate that even those three pumps have been operated at maximum speed only. We therefore threat all 11 pumps as ON-OFF pumps. Each time frame pumping rates mainly depend on 1) which pumps are switched on, and 2) the water drawdown in the corresponding wells. However, the range of water drawdown in Søndersø is fairly small compared to the range of the characteristics curves of the installed pumps. As consequence, we can say that average pumping rates are mainly conditioned by point 1) - which pump is running. The dependence on point 2) - the water drawdown underneath the running pumps - results in few percentage point-fluctuation of the pumping rates around the average. The plot in figure 12 shows several series having pump 1 either, switched on or off, all the time. At a first glance it appears that water level mainly goes down when pump 1 is switched off,

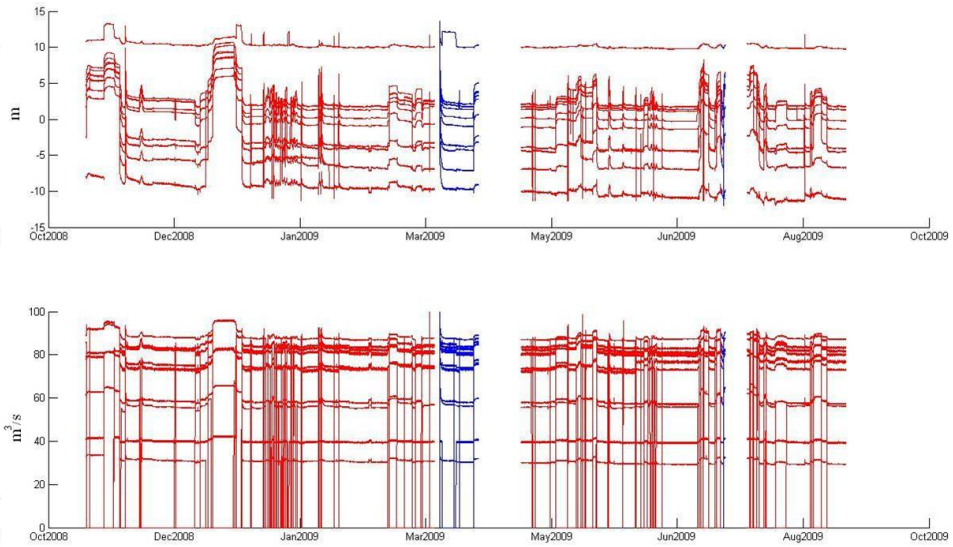


Figure 11: The five time series after data treatment

and vice versa when pump 1 is switched off. However individual patterns still appear rather irregular. Each series, the water level seems to alternate 1) relatively long periods with almost constant value, with 2) relatively short periods in which the value *jumps* to the next constant value. In other words the well field system iterates through different equilibriums, and those sudden transitions seem to happen when other some other pumps is switched ON/OFF. By defining *action* a variable $a_i \in \{0,1\}$ denoting the ON ($a = 1$) and OFF ($a = 0$) status of pump i, we define *decision* the vector of actions for all N pumps in the well field $a = (a_i, \ldots, a_N)$. In order to verify whether changes in decisions cause changes in equilibriums, data records are grouped in time series where decisions do not change. Datasets are created by grouping series with same decision; those data set are herein called *Stationary Decision Dataset* (SDD). For each possible decision, an SDD is populated. Although in Søndersø there are $2^{11} = 2048$ possible decisions, the available dataset only contains 93 different decisions. Among those, there are 58 decision with more than 20 records, 37 decisions with more than 1000 records, 6 decisions with more than 10,000 records. The only decision with more than 100,000 records (199,167) is $a = (1,1,1,1,1,1,1,1,1,1,1)$, hence all 11 pumps switched on.

4.2 Transitions, Oscillations and Noise

Once the data records have been gathered in Stationary Decision Datasets (SDDs), both water levels and pumping rates time series appear to draw regular shapes of the kind shown on Figure 13. Those shapes, in series of average duration (4200 minutes), seem to result from the overlapping of three mechanisms of decreasing ranges (see Figure 14)

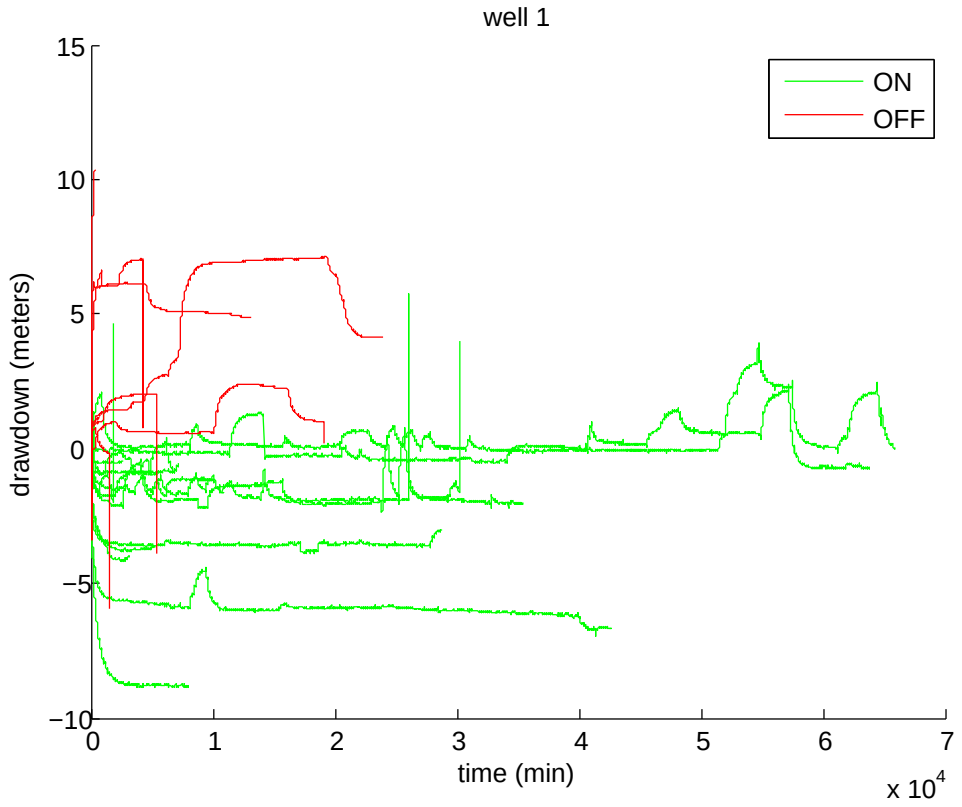


Figure 12: Drawdown levels at well 1. Each line is a period during which pump 1 was switched ON (ar OFF) all the time. Values are relative to the initial value.

1. The main mechanism is the **transition**; consisting of moving away from a starting value and to asymptotically approach a new quasi stationary value. We call *equilibriums* those stationary values.

2. The second mechanism consists of irregular **oscillations** around the equilibriums.

3. The third mechanism consists of **noises**, i.e. high frequency oscillations.

In terms of percentage of the average, transitions cause more variations in water levels than in pumping rates. In fact, as we said earlier, pumping rates always vary within relatively narrow ranges. Furthermore, as it can be easily observed on Figure 13, pumping rates converge faster than water levels. Qualitative observations on time series plots suggest that time series with similar pumping rates equilibriums also have similar water drawdown equilibriums. On the contrary, time series with different equilibriums in pumping rates also have equilibriums in water drawdown. Since pumping rates pumping rates values mainly depend on 1) which pumps are switched on, and weakly on 2) the water drawdown of the corresponding wells, it is natural to conclude that *transitions*

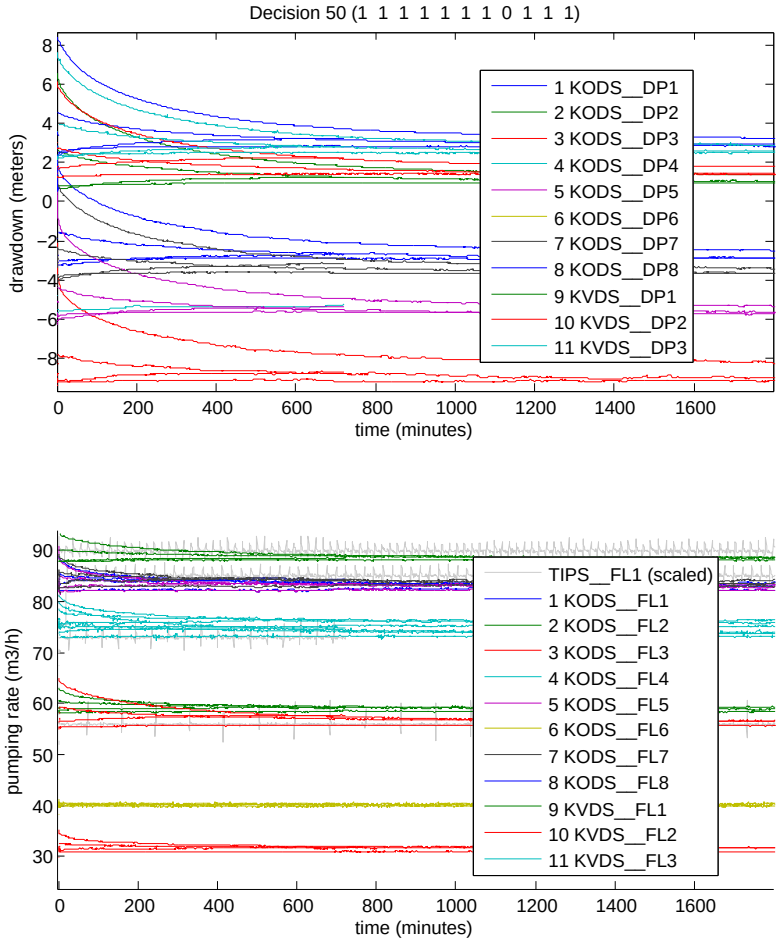


Figure 13: Drawdown levels and pumping rates time series in the SDD with decision $a = (11111110111)$. Grey curves are scaled series of total pumping rate of the South branch, q_t^s; labelled as TIPS_FL1.

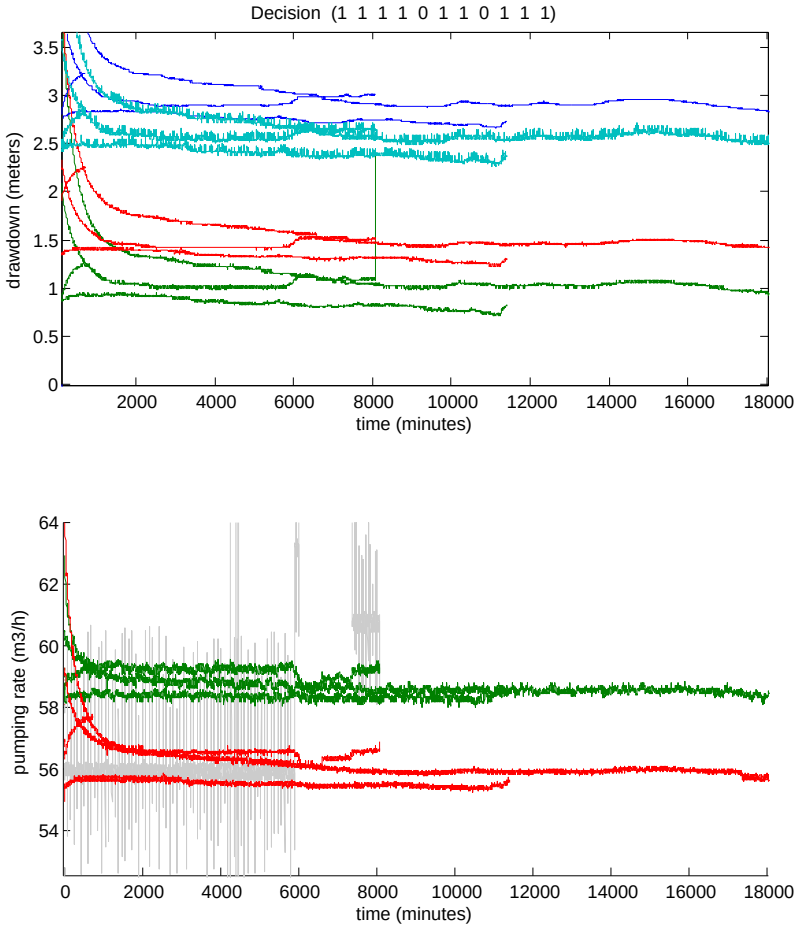


Figure 14: Transition (highlighted in green), oscillation (magenta), and noise (red frames) components of time variation of water level and pumping rates.

and equilibriums in drawdown equilibriums mainly depend on decisions.

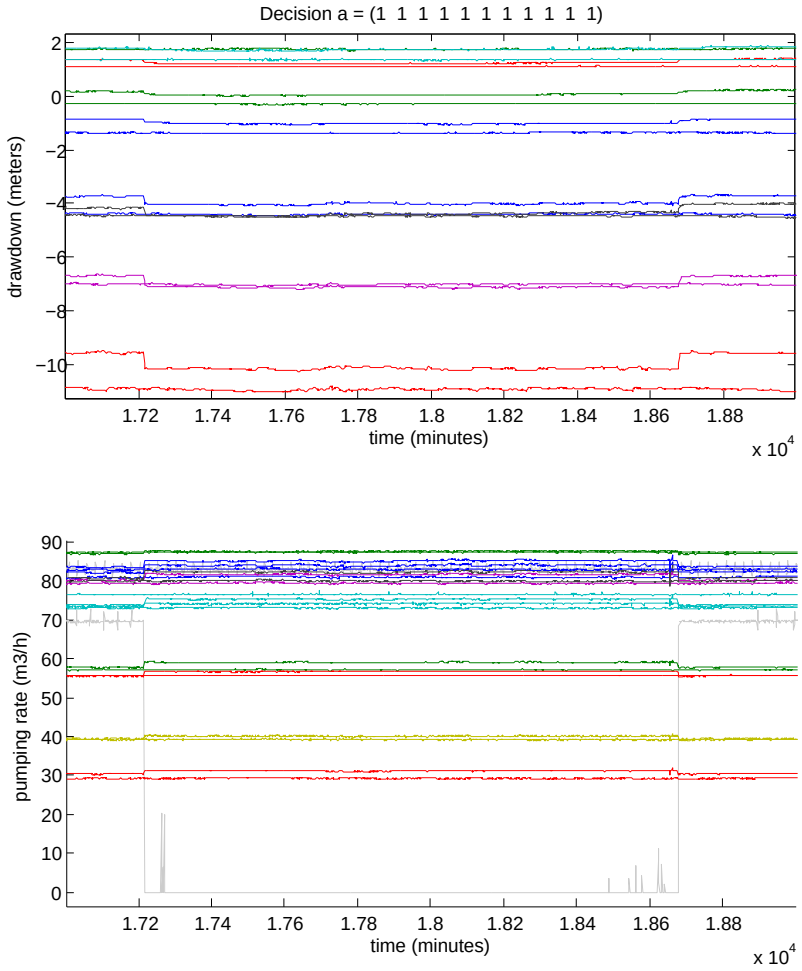


Figure 15: Drawdown levels. Each line refers to a well in a period during which decision is $a = (11111111111)$ all the time.

Once a decision is implemented and the system has completed the transition, oscillations are the main mechanisms for both pumping rates and water drawdown. such a mechanism can be triggered by two sources only:

1. time varying boundary conditions.

2. the variations in total pumpung rate q_s in the South branch of Søndersø, (series labelled as TIPS_FL1).

The contribution of this latter point can be observed in Figure 14 and in Figure 15, where discontinuous values of $q_s(t)$ correspond to discontinuous pumping rates and discontinuous water levels. The effect of variations in total pumping rate caused by q_s can be filtered out by re-creating the Stationary Decisions Datasets so that each SDD contains time series where both decisions and TIPS_FL1 pumping rates do not change. In fact,

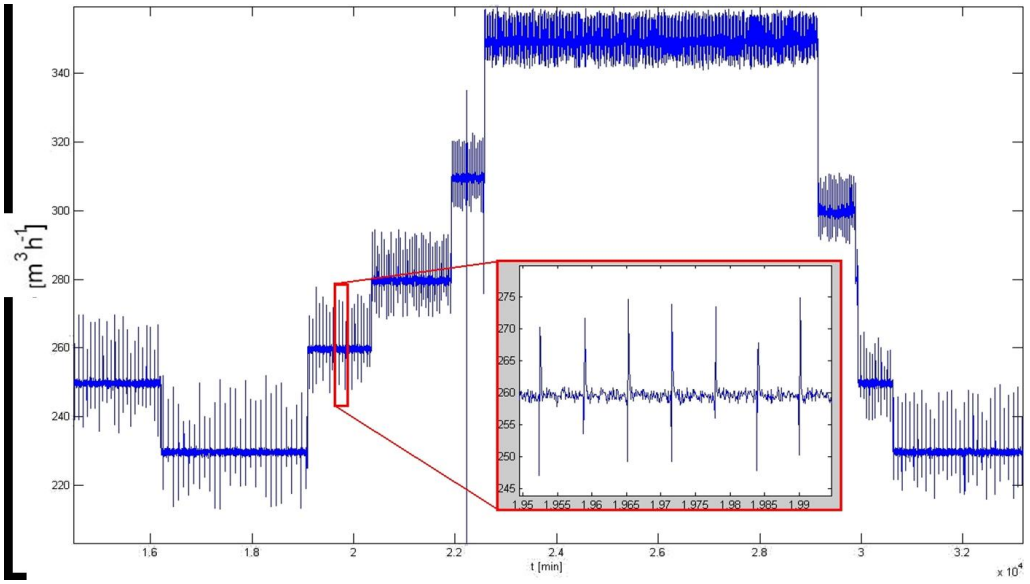


Figure 16: Stepwise pattern followed by the time series of the total pumping rate q_s of South Søndersø. Each jump corresponds to a change in pump settings.

even though records of individual pumping rates in south Søndersø are not available, it is still possible to precisely estimate when pumps setting changed in time. Time series of total pumping rate q_s follow a step-wise pattern, i.e. they move from constant value to constant values. This can be observed on the top chart in Figure 16. The second chart from the top of Figure 17 shows sampling frequencies of $q_s(t)$, the peacks indicate different pumps settings in south Søndersø. Whenever $q_s(t)$ value is on a *step*, it acts as a random variable whose distribution best estimate is always the normal distribution (third chart from the top of Figure 17). The fourth chart from the top of Figure 17 shows how time series have been grouped according to their distribution: series with *similar* standard deviations have been attributed to the same cluster. A total of 38 different pumps settings have been classified within the dataset at disposal (Figure 17 bottom). Following this, the dataset have been further broken down into SDDs containing series where both decisions and TIPS_FL1 pumping rates do not change. In sach SDD series oscillations should only be caused by time-dependent boundary conditions propagating across the well field.

4.3 Separating the three components

Next and last task described in this subsection is the attempt to separate the three components, (transitions, oscillation and noise), from a given time series. The followed approach is based on the observations that transitions seem to follow patterns similar to stable linear systems. During a transition, in fact, the water level $h_i(t)$ at well i moves away from the starting level h_i^0 and progressively approaches the stationary level h_i^∞.

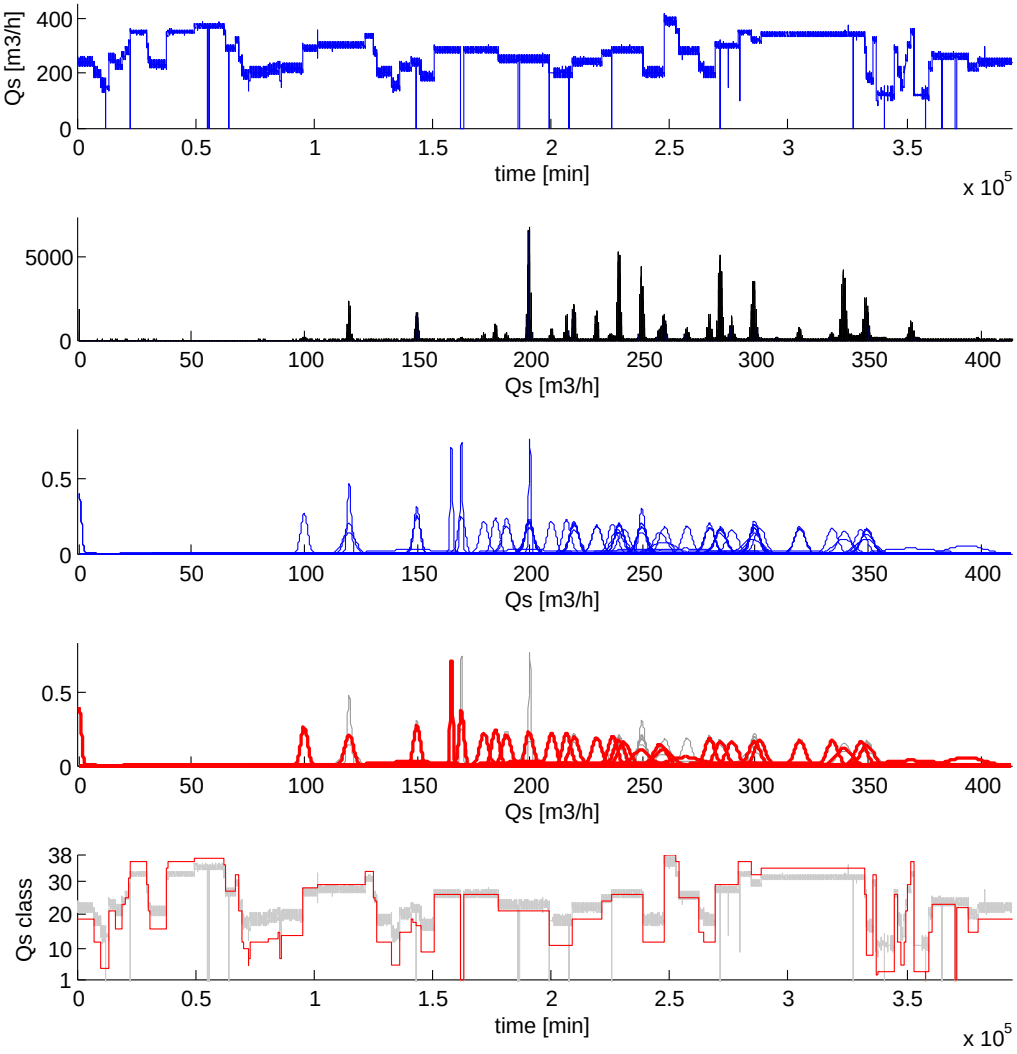


Figure 17: Clustering of total pumpung rate q_s in the South branch of Søndersø, (series labelled as TIPS_FL1).

Judging from the plots (Figures 15 and 13), the way this happens looks similar to the dynamics of a stable ($\alpha < 0$) linear system

$$\frac{dh_i}{dt} = \alpha h(t) - \alpha h_i^{\infty} \tag{4}$$

integrated

$$h_i(t) = \left(h_i^0 - h_i^{\infty}\right) e^{\alpha(t-t_0)} + h_i^{\infty}$$

So the idea is to calibrate a linear system for each time series of each DSS, and to subtract the obtained curve from the original data; the residual are the oscillations.

First order derivatives dh_i/dt, can be calculated by setting an appropriate sampling time interval Δt to subsample the time series and approximate with $dh_i/dt \simeq (h_i(t + \Delta t) - h_i(t))/\Delta t$, see section 5. Such an approximation is accurate as long as the noise components of the time series, which is a potentially striking element of disturbance, is reduced. As discussed in section 1, noises are caused by measuring errors. They can be effectively removed from the data by employing any smoothing technique, in this case *spline* have been used (see for instance de Boor (1978)). Figures 18, 19 illustrate the how smoothing and linear systems are employed to separate the three components, transition, oscillation and noise, from a time series taken from a SDD. The raw data are smoothed and noise component is obtained as difference between raw and smoothed data. The smoothed data are also used to calibrate the linear model, which is then simulated and the resulting curve is subtracted from the smoothed data. The result is the oscillation time series, this time cause by boundary conditions only. By putting together oscillations time series, it is possible to etimate the water levels at each well if no pumping was done. An example is on Figure 20, showing that water level in one month, betwee November 2008 and December 2008 has dept half meter. This explains why pumping rates are diffenent in scenarios having the same decision applied. They are different because the level of the aquifer is different; as mentioned earlier, pumping rates depend on decisions and water drawdown.

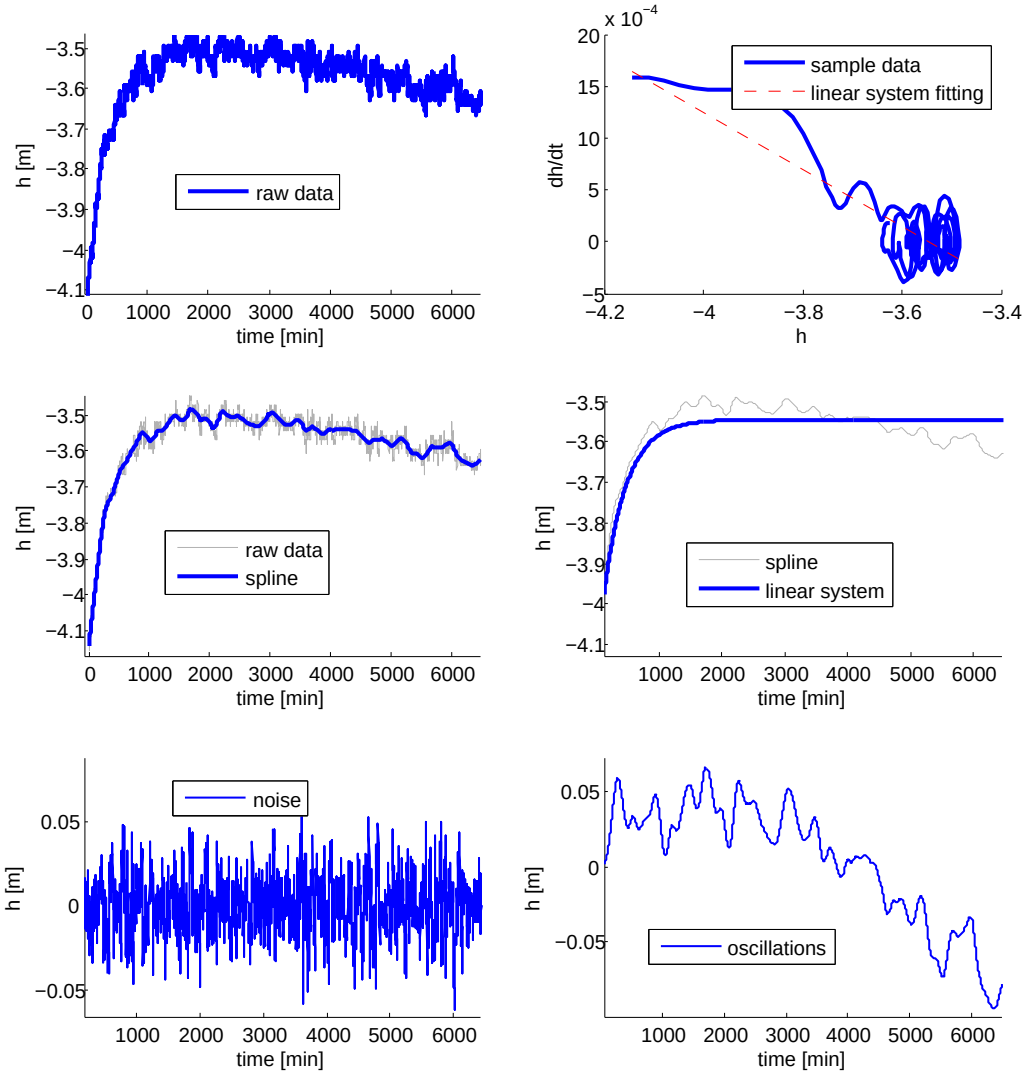


Figure 18: Separation of Transition, Oscillation and Noise components of time series.

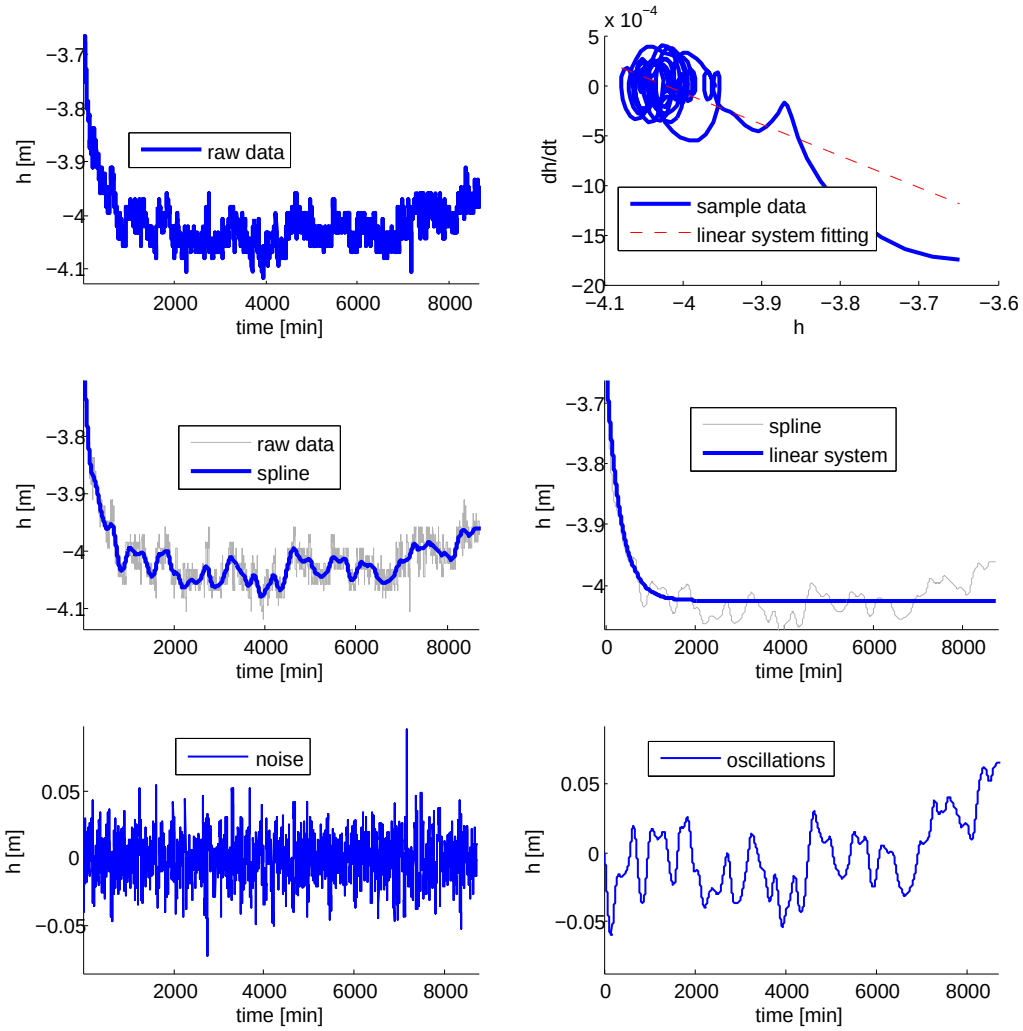


Figure 19: Separation of Transition, Oscillation and Noise components of time series.

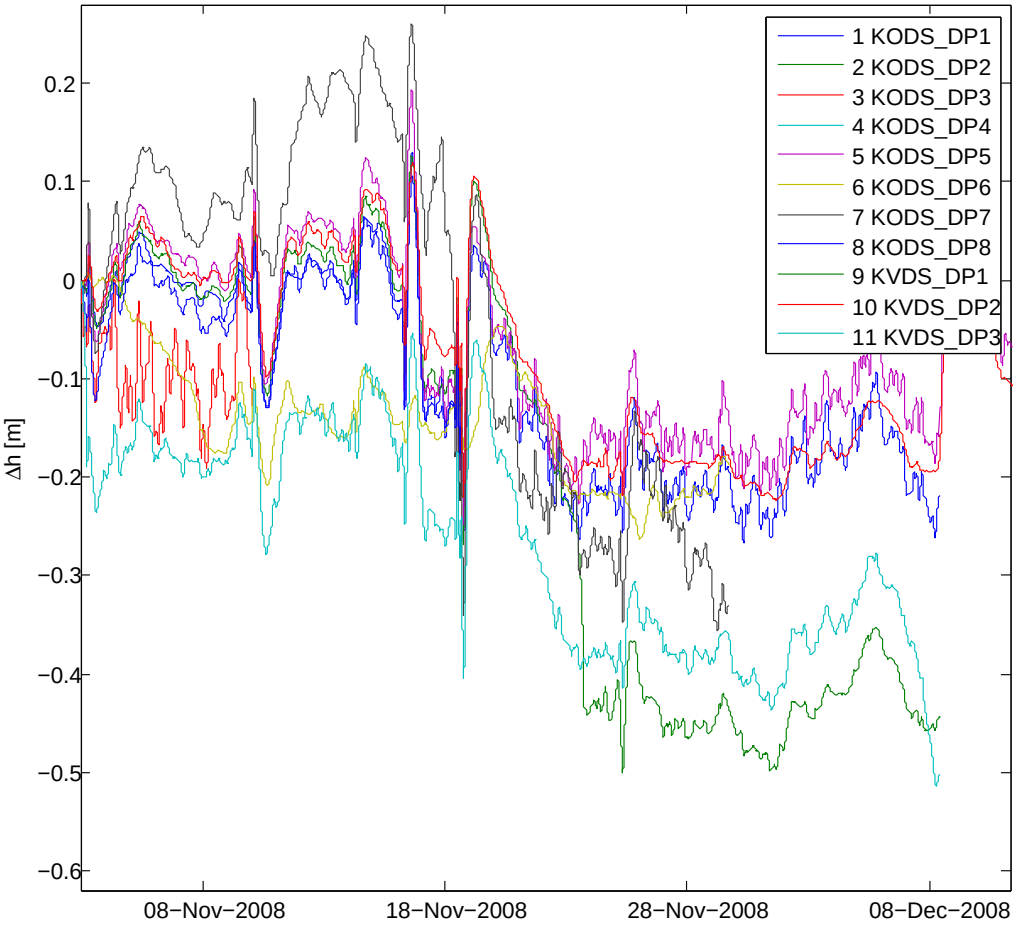


Figure 20: Estimation of the water level of the aquifer at wells points, if no pumping was performed

4.4 A linear representation

During a transition, the water level $h_i(t)$ at well i moves away from the starting level h_i^0 and progressively approaches the stationary level h_i^∞. Judging from the plots (Figures 15 and 13), the way this happens looks similar to the dynamics of a stable ($\alpha < 0$) linear system

$$\frac{dh_i}{dt} = \alpha h(t) - \alpha h_i^\infty \tag{5}$$

integrated

$$h_i(t) = \left(h_i^0 - h_i^\infty\right) e^{\alpha(t-t_0)} + h_i^\infty$$

This sections verifies whether the observed water level $h_i(t)$ at well i can be fitted within a linear model. In case of good fitting, this is likely to bring important insights regarding the structure of the final grey-box model. The type of linear model herein investigate has form

$$\frac{dh_i}{dt} = \left[\alpha_0 + \alpha_1 \cdot q_1(t) + \ldots + \alpha_N \cdot q_N(t)\right] \cdot h_i(t) \quad + \quad \beta_0 + \beta_1 \cdot q_1(t) + \ldots + \beta_N \cdot q_N(t) \tag{6}$$

Where coefficients $\alpha_k = \alpha_k(i,a)$,$\beta_k = \beta_k(i,a)$ for all $k = 0, \ldots, N$, are constants as depending on the well i and the decision a, which are both constants. From now on, unless specified, the well i and decision a are considered fixed;

First order derivatives dh_i/dt, are calculated the way as described in section 4. We denote with M the number of time series within the dataset of scenario having exclusively decision a applied on the well field. The most immediate way to calibrate a model 6 is to solve the system of linear equations $A \cdot x = b$, where

$$A = \begin{bmatrix}
1 & h_i^1(0) \cdot q_1^1(0) & \cdots & h_i^1(0) \cdot q_N^1(0) & 1 & q_1^1(0) & \cdots & q_N^1(0) \\
1 & h_i^1(\Delta t) \cdot q_1^1(\Delta t) & \cdots & h_i^1(\Delta t) \cdot q_N^1(\Delta t) & 1 & q_1^1(\Delta t) & \cdots & q_N^1(\Delta t) \\
1 & h_i^1(2 \cdot \Delta t) \cdot q_1^1(2 \cdot \Delta t) & \cdots & h_i^1(2 \cdot \Delta t) \cdot q_N^1(2 \cdot \Delta t) & 1 & q_1^1(2 \cdot \Delta t) & \cdots & q_N^1(2 \cdot \Delta t) \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & h_i^2(0) \cdot q_1^2(0) & \cdots & h_i^2(0) \cdot q_N^2(0) & 1 & q_1^2(0) & \cdots & q_N^2(0) \\
1 & h_i^2(\Delta t) \cdot q_1^2(\Delta t) & \cdots & h_i^2(\Delta t) \cdot q_N^2(\Delta t) & 1 & q_1^2(\Delta t) & \cdots & q_N^2(\Delta t) \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & h_i^M(0) \cdot q_1^M(0) & \cdots & h_i^M(0) \cdot q_N^M(0) & 1 & q_1^M(0) & \cdots & q_N^M(0) \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{bmatrix}$$

$$b = \begin{bmatrix} \frac{dh_i^1(0)}{dt} & \frac{dh_i^1(\Delta t)}{dt} & \frac{dh_i^1(2 \cdot \Delta t)}{dt} & \cdots & \frac{dh_i^2(0)}{dt} & \frac{dh_i^2(\Delta t)}{dt} & \cdots & \frac{dh_i^M(0)}{dt} & \cdots \end{bmatrix}^T$$

$$x = \begin{bmatrix} \alpha_0 & \alpha_1 & \cdots & \alpha_N & \beta_0 & \beta_1 & \cdots & \beta_N \end{bmatrix}^T$$

This over determined system can be solved using linear least squares.

$$x = (A^T A)^{-1} A^T b \tag{7}$$

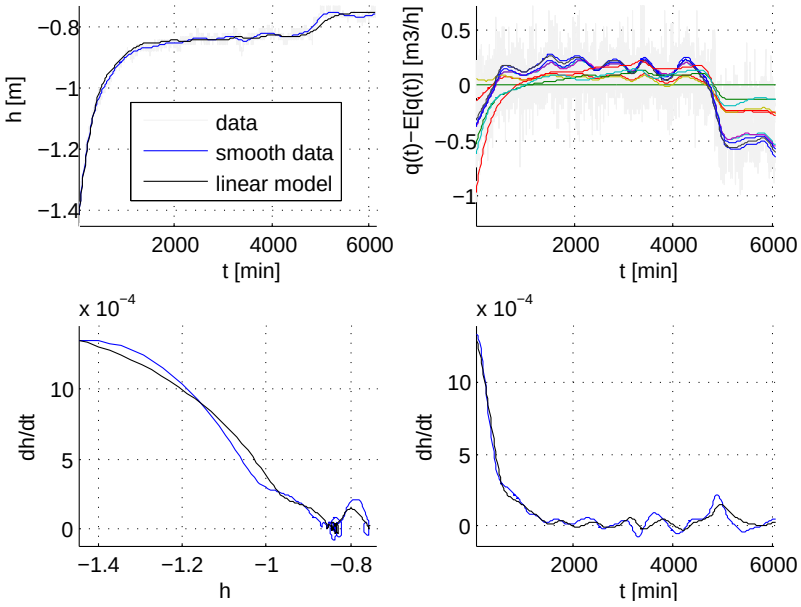


Figure 21: Linear model testing for drawdown simulation. Well $i = 8$, decision $a = (10111111111)$, $R^p = .986$, $R^s = .996$.

Although such a procedure is straightforward, it is important to flag that this is the way to calibrate a *predictor* rather than a *simulator*. The difference is that a predictor requires a measure $h_i(t)$ in order to estimate $\widehat{h}_i(t + \Delta t)$:

$$\frac{d\widehat{h}_i}{dt} = [\alpha_0 + \alpha_1 \cdot q_1(t) + \ldots + \alpha_N \cdot q_1(t)] \cdot h_i(t) \quad + \quad \beta_0 + \beta_1 \cdot q_1(t) + \ldots + \beta_N \cdot q_1(t)$$

$$\widehat{h}_i(t + \Delta t) = h_i(t) + \frac{d\widehat{h}_i}{dt} \cdot \Delta t$$

A simulator instead, requires the measurement of the water level at the beginning of the simulation only, i.e. $\widehat{h}_i(0) := h_i(0))$. The value $h_i(0))$ is utilised to calculate $\widehat{h}_i(\Delta t)$. The calculated value $\widehat{h}_i(\Delta t)$ is then utilised to calculate $\widehat{h}_i(2 \cdot \Delta t)$ (a predictor instead would require the actual $h_i(\Delta t)$ to be measured and passed as input. Ideally one would like to calibrate the parameters of model 6 as simulator, however this requires to solve a much more complex system of non-linear equations, whose solution is non as immediate as for the linear least squares. The results are encouraging, as in most of the cases simulations are able to follow the time series with good accuracy (Figures 21,24 and 25). However no all simulations are successful; Figures 22 and 23 show two examples of unstable models where the error build up causes the model to drift away from the reality. A linear system of type of equation 5 is stable if and only if $\alpha < 0$. The system of form 6 is stable when

$$\alpha_0^a + \alpha_1^a \cdot q_1(t) + \ldots + \alpha_N^a \cdot q_1(t) < 0, \forall t \tag{8}$$

and this may happen not for all time t, or even never.

One way to act could be to solve the system of linear equations $A \cdot x = b$ plus as many conditions of type 8 as the records of the sub-sampled time series. The resulting problem

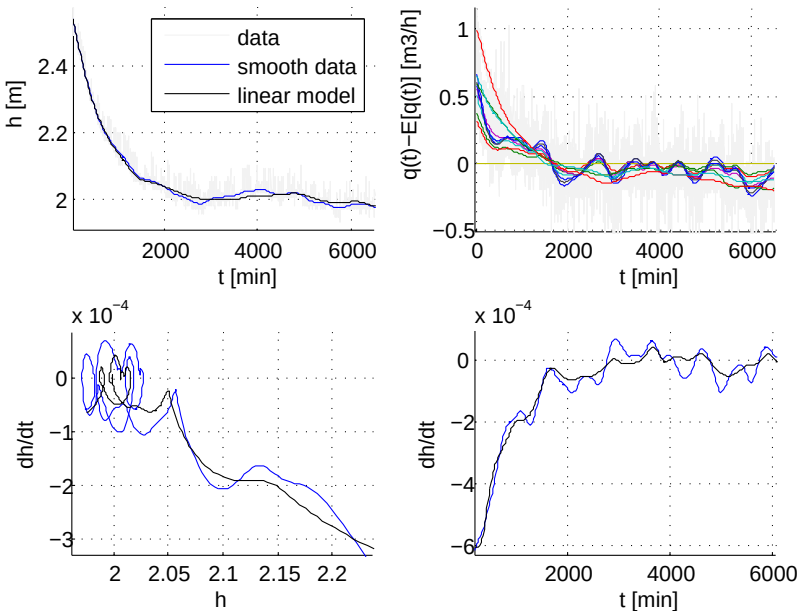


Figure 22: Linear model testing for drawdown simulation. Well $i = 11$, decision $a = (11111011111)$, $R^p = .977$, $R^s = .997$.

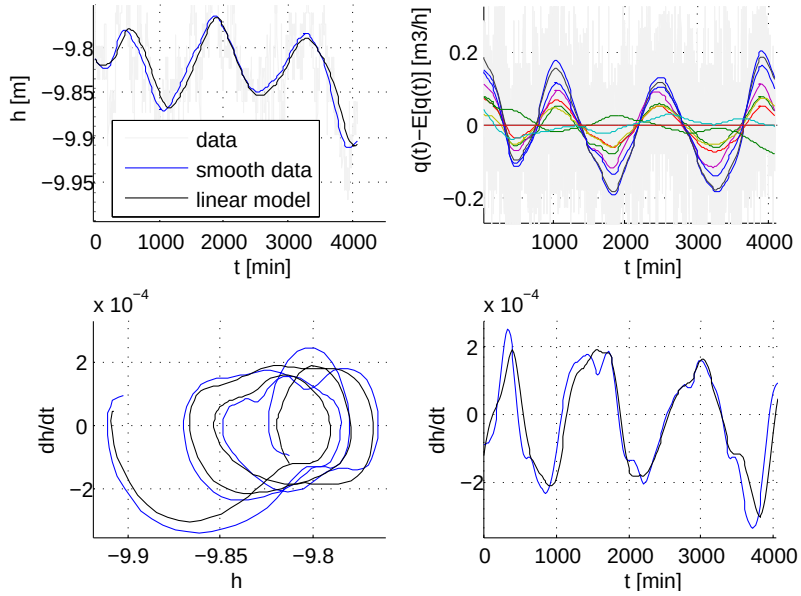


Figure 23: Linear model testing for drawdown simulation. Well $i = 3$, decision $a = (11101111101)$, $R^p = .954$, $R^s = .962$.

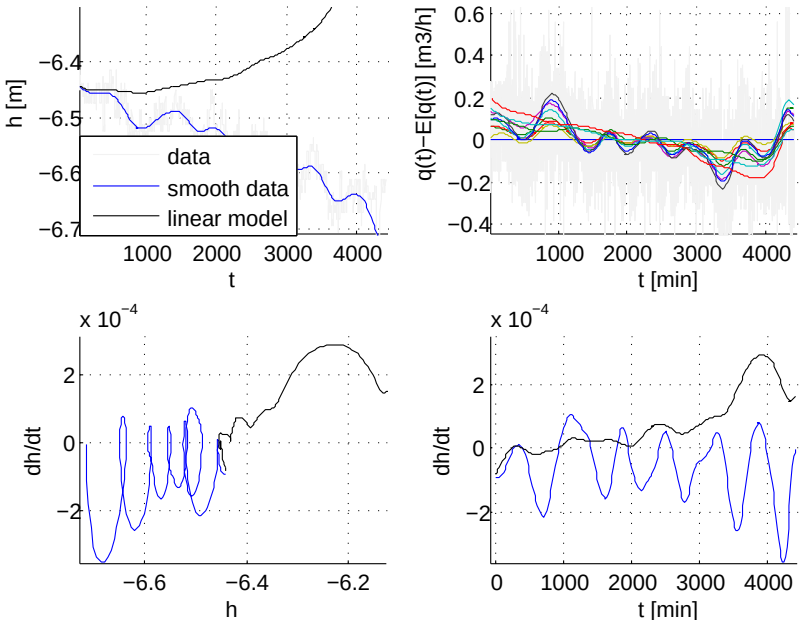


Figure 24: Linear model testing for drawdown simulation. Well $i = 5$, decision $a = (01111111111)$, $R^p = .977$, $R^s = -.912$.

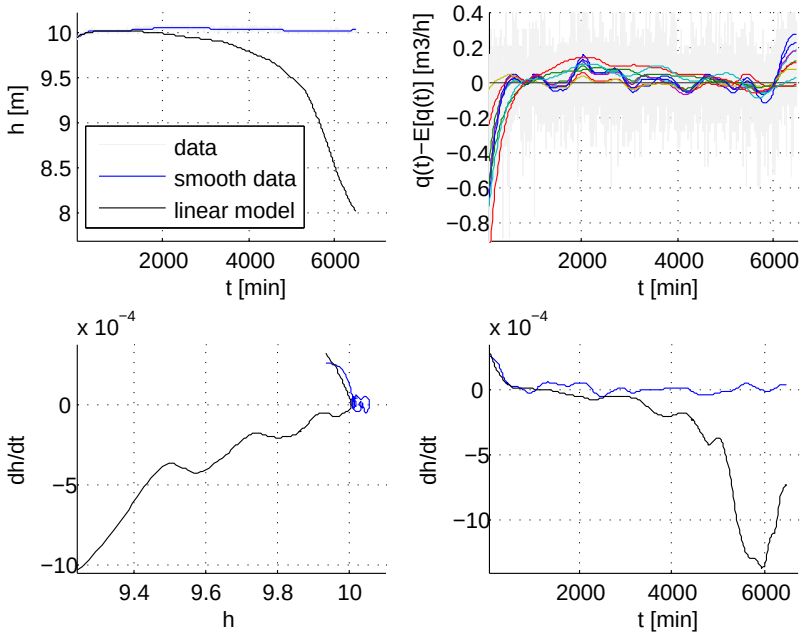


Figure 25: Linear model testing for drawdown simulation. Well $i = 6$, decision $a = (11111101111)$, $R^p = .968$, $R^s = .194$.

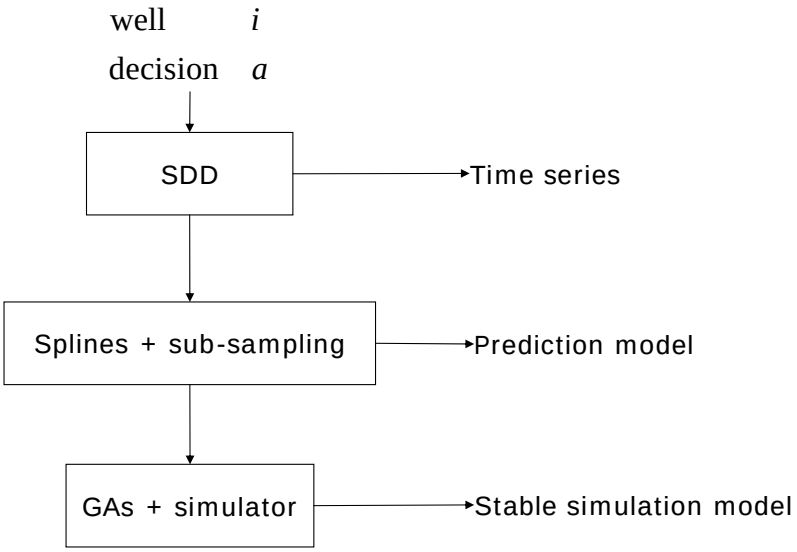


Figure 26: The adopted procedure to produce stable and reliable simulation models.

is a *quadratic programming* problem. Such a problem can be tackled using for instance the subspace trust-region method (Coleman & Li (see 1992)). However even this would produce another predictor, whereas a simulator is still needed.

The approach herein adopted to produce stable and reliable simulators consist in two steps, see scheme in Figure 26. Firstly a predictor is calibrated using least square, as explained. Secondly the predictor is improved and transformed into a simulator by using Genetic Algorithms. Genetic Algorithms (GAs) are adaptive heuristic search algorithms. First pioneered by John Holland in the 60s, Genetic Algorithms have been widely studied, experimented and applied in many fields of engineering, see Davis & Mitchell (1991). GAs can be considered as a black-box approach to the optimization, as they work regardless the type of the problem. In this framework, GAs work iteratively. Each iteration a certain number of possible solutions, i.e. sets of candidate parameters, $\alpha_0, \alpha_1, \ldots, \alpha_N, \beta_0, \beta_1, \ldots, \beta_N$ are generated by the GA. Each set of parameters is passed to a simulator returning the M simulations $\widehat{h}_i^1(0), \widehat{h}_i^1(\Delta t), \widehat{h}_i^1(2 \cdot \Delta t), \ldots, \widehat{h}_i^2(0), \widehat{h}_i^2(\Delta t) \ldots, \widehat{h}_i^M(0), \ldots$ The simulations are then compared with the time series and the level of fitness is calculated as total square error:

$$D = \sum_{j=1}^{M} \overbrace{(\widehat{h}_i^j(0) - h_i^j(0))^2}^{=0 \text{ initial value}} + (\widehat{h}_i^j(\Delta t) - h_i^j(\Delta t))^2 + (\widehat{h}_i^j(2\Delta t) - h_i^j(2\Delta t))^2 + \cdots$$

Once evaluated, solutions are ranked according to their fitness and utilised to generate new improved solutions for the next iteration. Iterations carry on until convergence is reached (i.e. until some stopping criterion is fulfilled).

As said above, GAs are herein employed to improve the model resulting from the least squares. The parameters of such a model are in fact passed to the GA optimizer as initial

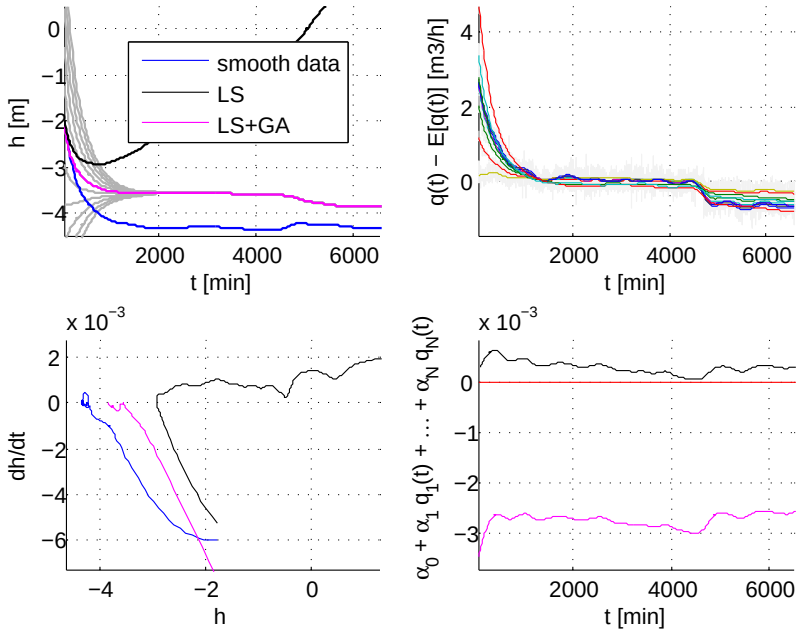


Figure 27: Stable linear models for drawdown simulation. Well $i = 10$, decision $a = (11111111111)$

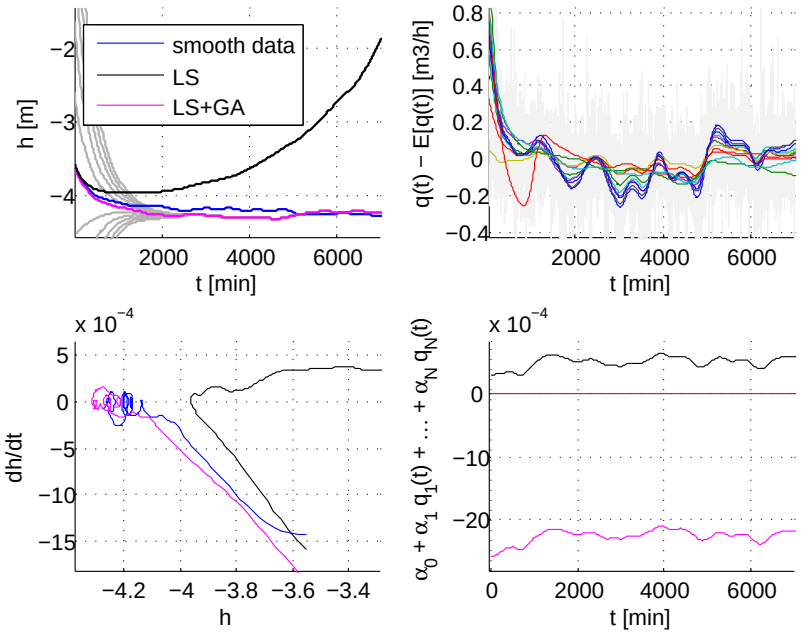


Figure 28: Stable linear models for drawdown simulation. Well $i = 10$, decision $a = (11111111111)$

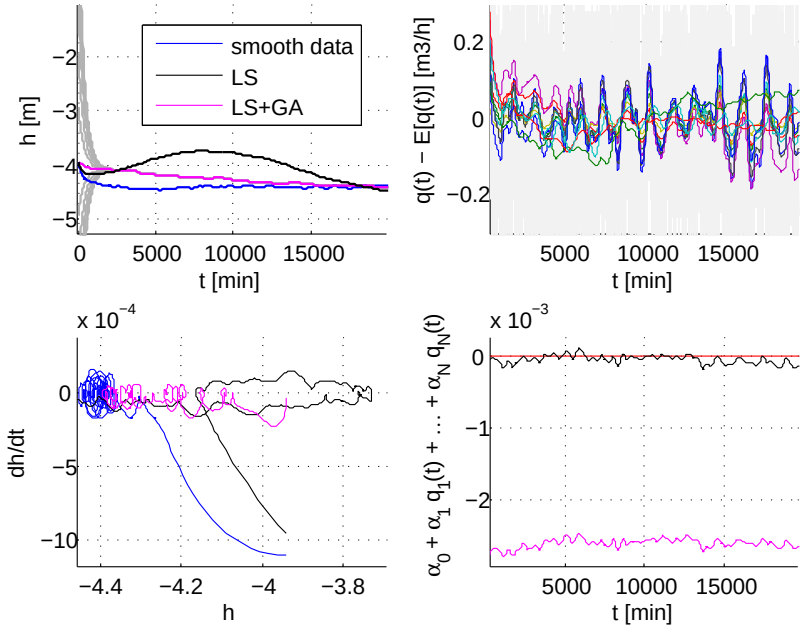


Figure 29: Stable linear models for drawdown simulation. Well $i = 10$, decision $a = (11111111111)$

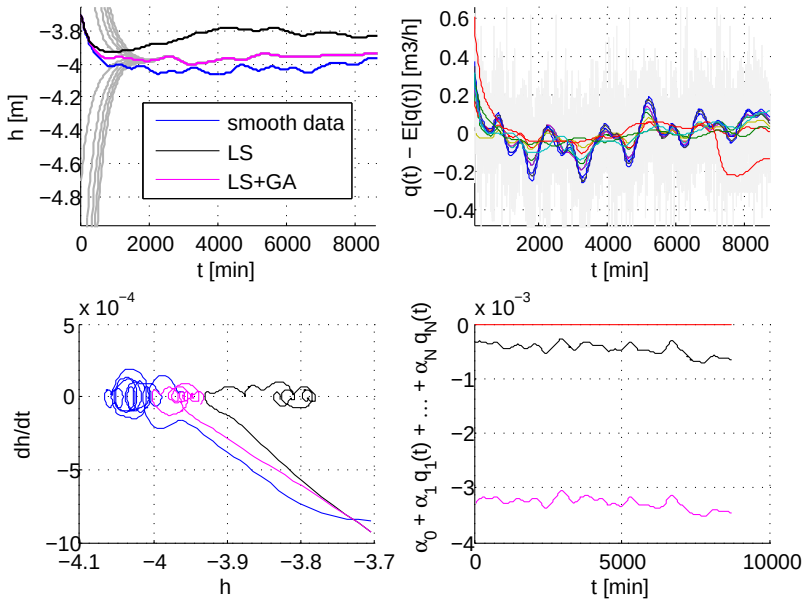


Figure 30: Stable linear models for drawdown simulation. Well $i = 10$, decision $a = (11111111111)$

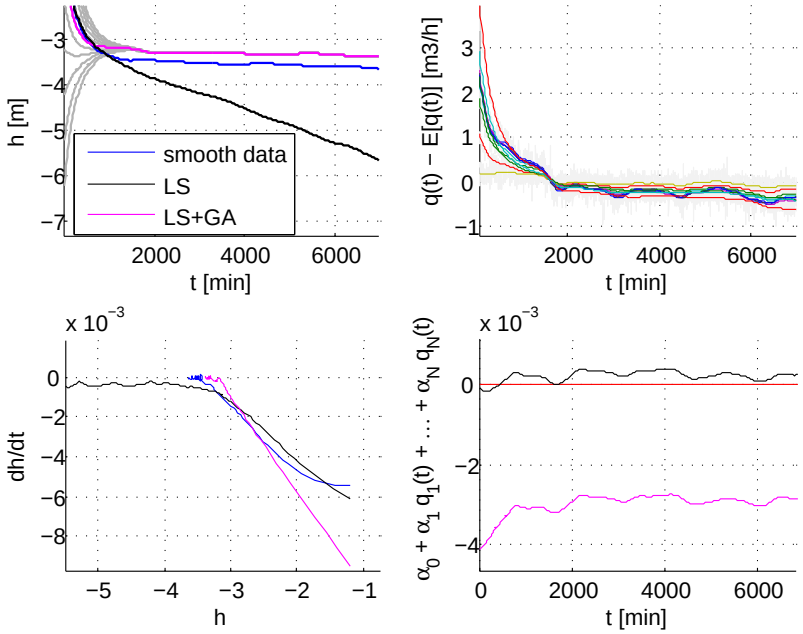


Figure 31: Stable linear models for drawdown simulation. Well $i = 10$, decision $a = (11111111111)$

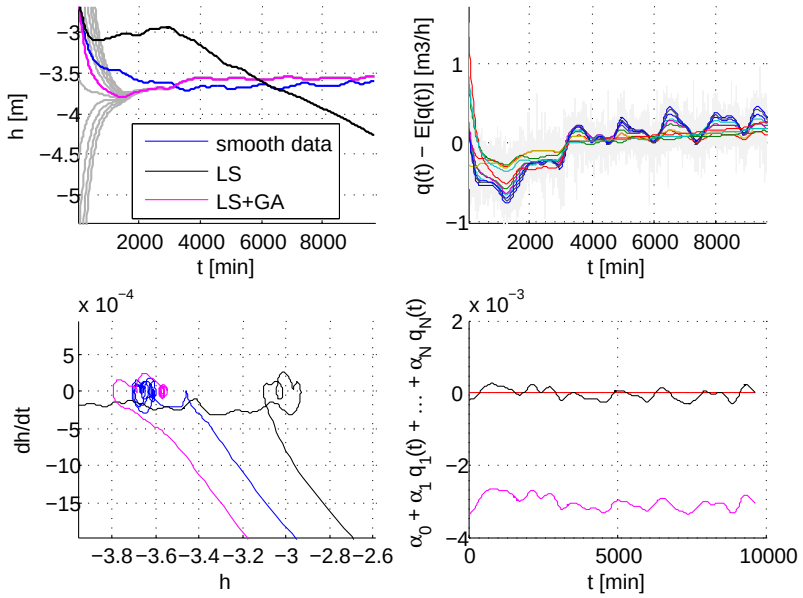


Figure 32: Stable linear models for drawdown simulation. Well $i = 10$, decision $a = (11111111111)$

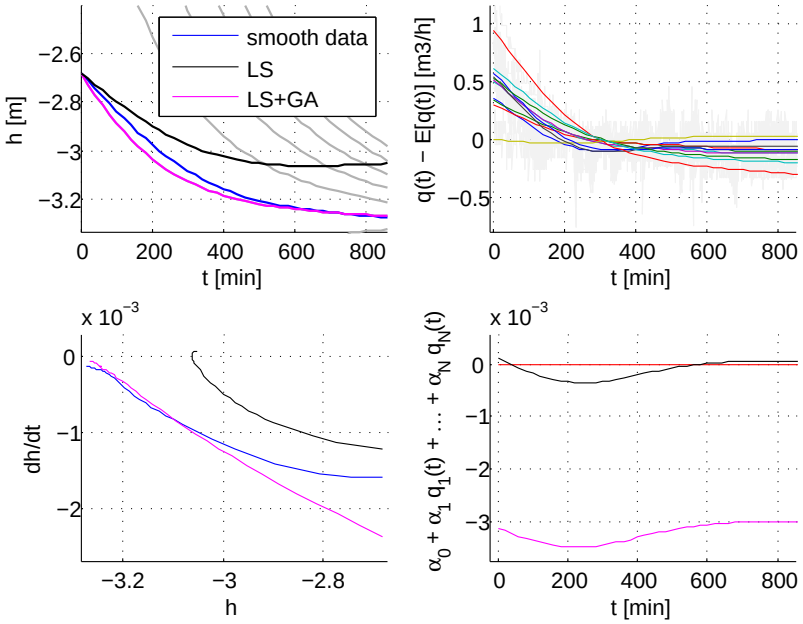


Figure 33: Stable linear models for drawdown simulation. Well $i = 10$, decision $a = (11111111111)$

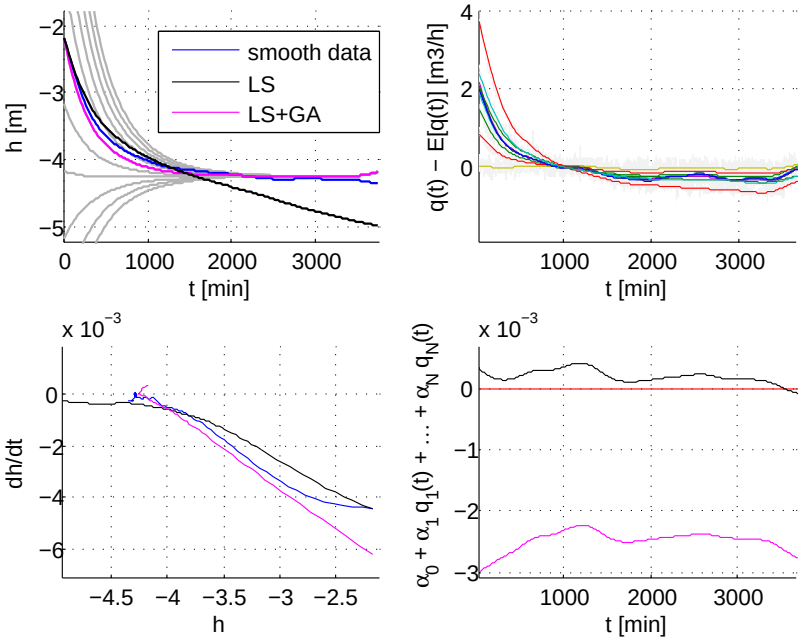


Figure 34: Stable linear models for drawdown simulation. Well $i = 10$, decision $a = (11111111111)$

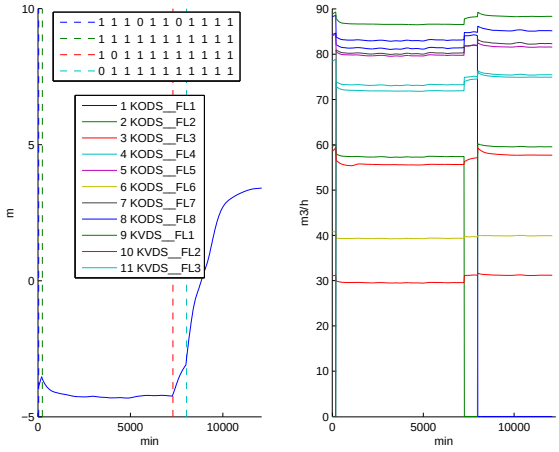


Figure 35: Simulations of water drawdown of well $i = 1$ with changing decision.

point to seed the algorithm. Least square optimization results are also exploited to set the boundary of the GA search space. Consider again the least square formula 7

$$x = (A^T A)^{-1} A^T b$$

in a least squares calculation, or in linear regression, the variance on the jth parameter x_j, denoted $\text{var}(x_j)$, is usually estimated with

$$\text{var}(x_j) = \sigma^2 \left(\left[A^T A \right]^{-1} \right)_{jj} \approx \frac{S}{n - N} \left(\left[A^T A \right]^{-1} \right)_{jj}$$

where σ^2 is the true residual variance, whereas S is the sum of square errors, and n is the number of rows of A, namely the total number of records utilised for calibration. Once mean and variance of parameters estimations are calculated, confidence intervals can be determined. For each parameter $\alpha_0, \alpha_1, \ldots, \alpha_N, \beta_0, \beta_1, \ldots, \beta_N$, the GA search space is constrained to vary within its 95% confidence interval.

Some results are in Figures 27-34, showing some series where the linear model model for well $i = 1$ decision $a = (111111111)$ was unstable and made stable by the GA optimization. The bottom-right charts of the aforementioned figures show that for all parameter estimations the GA enhanced model succeeded to fulfill stability condition 8. The grey plots on top-left charts are simulations where the initial conditions are perturbed; the stable system leads the water drawdown estimation to the equilibrium level.

Having models calibrated for different decisions allows for simulations of pumping scenarios with variable decisions. Every time the decision changes the model switches to the parameters corresponding to the new decision. Some examples of simulations are in Figures 35-38.

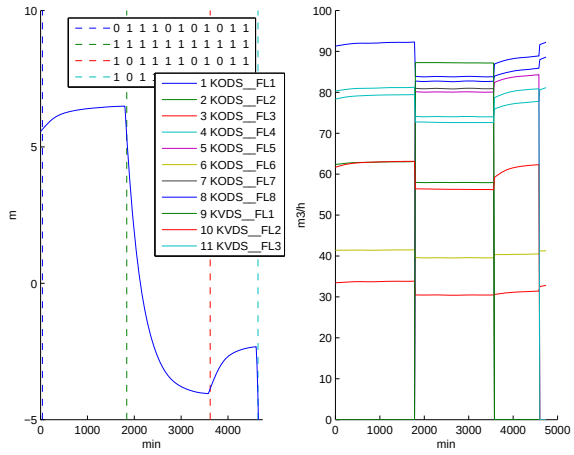


Figure 36: Simulations of water drawdown of well $i = 1$ with changing decision.

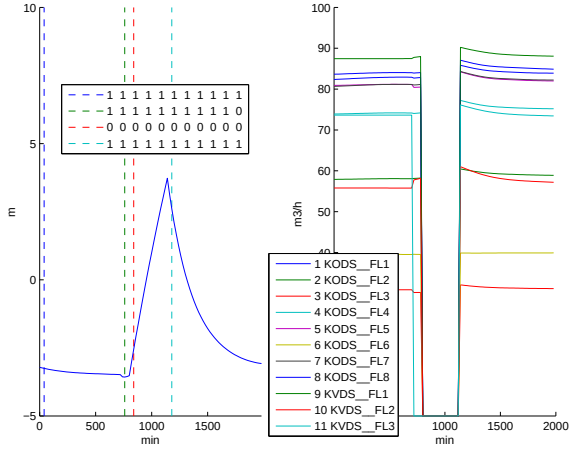


Figure 37: Simulations of water drawdown of well $i = 1$ with changing decision.

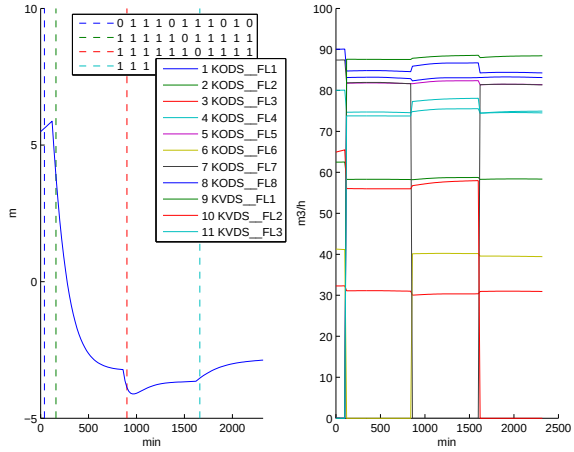


Figure 38: Simulations of water drawdown of well $i = 1$ with changing decision.

5 Sampling the treated datasets

Based on the treated data set, an appropriate intervals and sampling periods can be determined for further investigation of the data. This is a fundamental step towards getting an adequate time series for developing a model structure that can be utilized for simulations and predictions of drawdowns in the well field. A sufficient time series for the structure must grasp the main matters in the process, essentially the dynamics of the system and instantanious interactions affecting the model output.

As previously mentioned in Section 4, for changes in decisions the consequent drawdown is changed as well and reaches a equilibrium level as long as the strategy is not changed again within several hours (SDD). This can be seen in Figure 13 for all wells on the East and West. However, to analyze the drawdown in each of the wells, where one or more drawdowns for an arbitrary decision for several hours is available, the traces have been separated as displayed in Figures 39 and 40 for switching on or off, respectively, i.e. each plot is assigned to a well, which has available data in the data set from the Søndersø well field, showing all measured drawdowns (in black and labeled to the left) for corresponding pumping rate (in grey and labeled to the right).

Figure 39 shows the water level responding to the event when a pump is switched on. This is taken from the first treated time series, thus, only 9 out of 11 wells are shown since the majority of observations for two of the well drawdowns, namely KODS_4 and KVDS_1, is corrupted. As expected does the water level sink when water is withdrawn from the wells and for most wells, the drawdown approaches a steady-state. The water level appears to depend on the elevation when the pump rate was turned on. Although the drawdown seems to be quite the same for the individual well, it is depending on the decision.

In many cases the drawdown is twofold; first there is a large drop in the first minutes, followed by more smooth drawdown approaching steady-state the following hours. The significant drop in the first minutes is due to the direct interaction between the measured pumping rate and corresponding drawdown level; adrupt changes in the pumping cause the waterlevel in the well to drop accordingly. However, the waterlevel in the well field is first detected when the water drawdown in the well has reached an equilibrium regarding the decision. Hence, the change in the well field correspond to the exponential decay for switching pump on, or exponential rise for pump being turned off.

This indicates that there are two time constants in the system. The first one is observed in the large leap in the first minutes from decision. The second time constant can be detected from the individual well; the time duration from dicision to the equilibrium is similar for all traces of water drawdown in the well, despite the starting value for the water drawdown. From the plots it can be seen that the second time constant is approximately 4 hours.

Figure 40 shows the water drawdown in the wells when corresponding pump is switched off. As seen from the plots very few track records for reasonable long period are available for a single well being turned off, i.e. the decision is changes within short time. For both

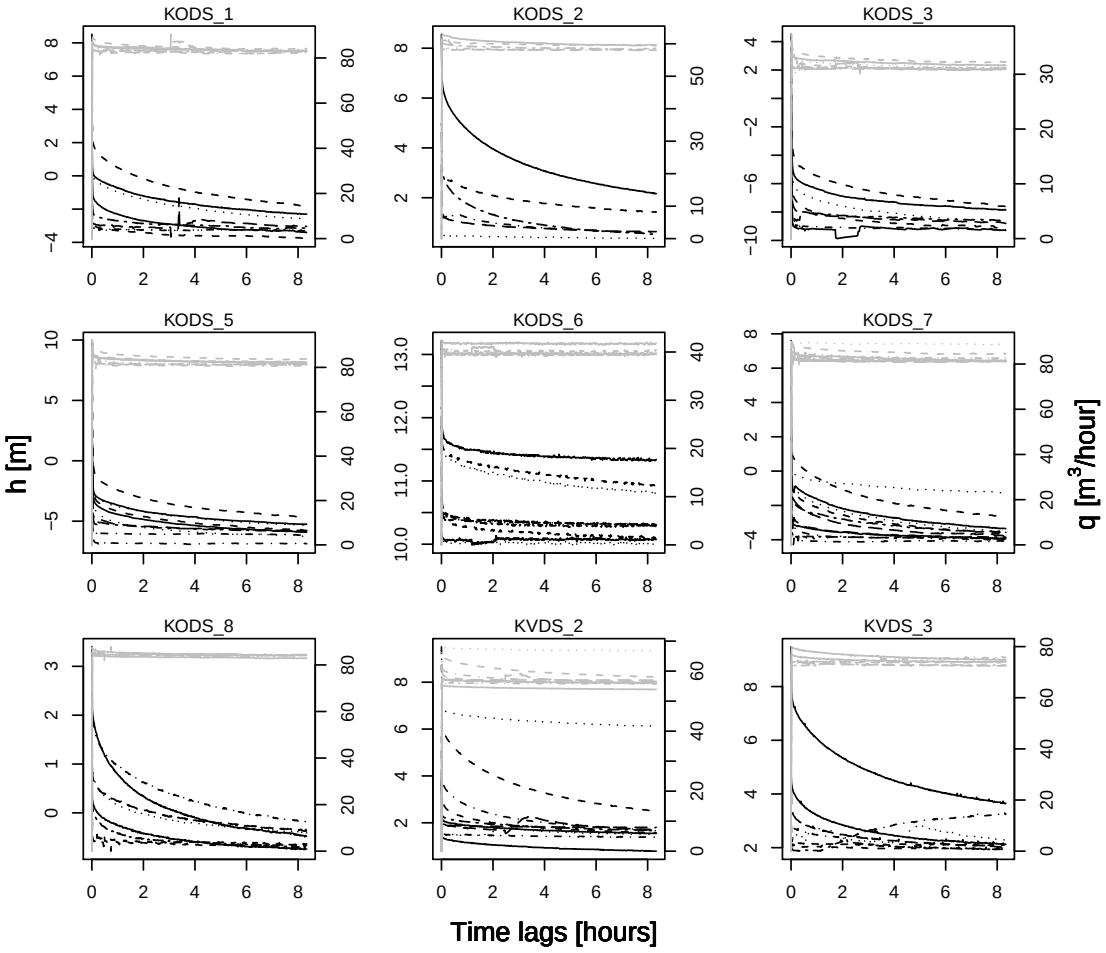


Figure 39: Step response for the individual drawdown in each well. Drawdown when corresponding pumping rate is turned on.

KODS_3 and KVDS_3 only one track spans the time range defined to detect the time constant. For the first minutes from decision the same adrupted change occur as in Figure 39, but the following increasing exponential seems to approach the equilibrium slower. This is not in general the case; e.g. wells KODS_2, KODS_8 and KVDS_3 display similar response as the average drawdown for switching on a well (Figure 39). Regarding the remaining wells it appears as the time constant is longer, but the one minute resolution of the data is not able to give a clear image of the first constant, which then affects the estimation of the second time constant. In theory would we expect the same time constant for all well since pumping from the same aquifer and because of the low resolution in the observations we can with assume the second time constant is the same for discharge and recharge of the aquifer.

The drawdown and the infiltration are in continuous time, which means that what appears in Figures 39 and 40 are just linear interpolations between consecutive discretely observed data points for the water levels in the wells. The drop in the first minute indi-

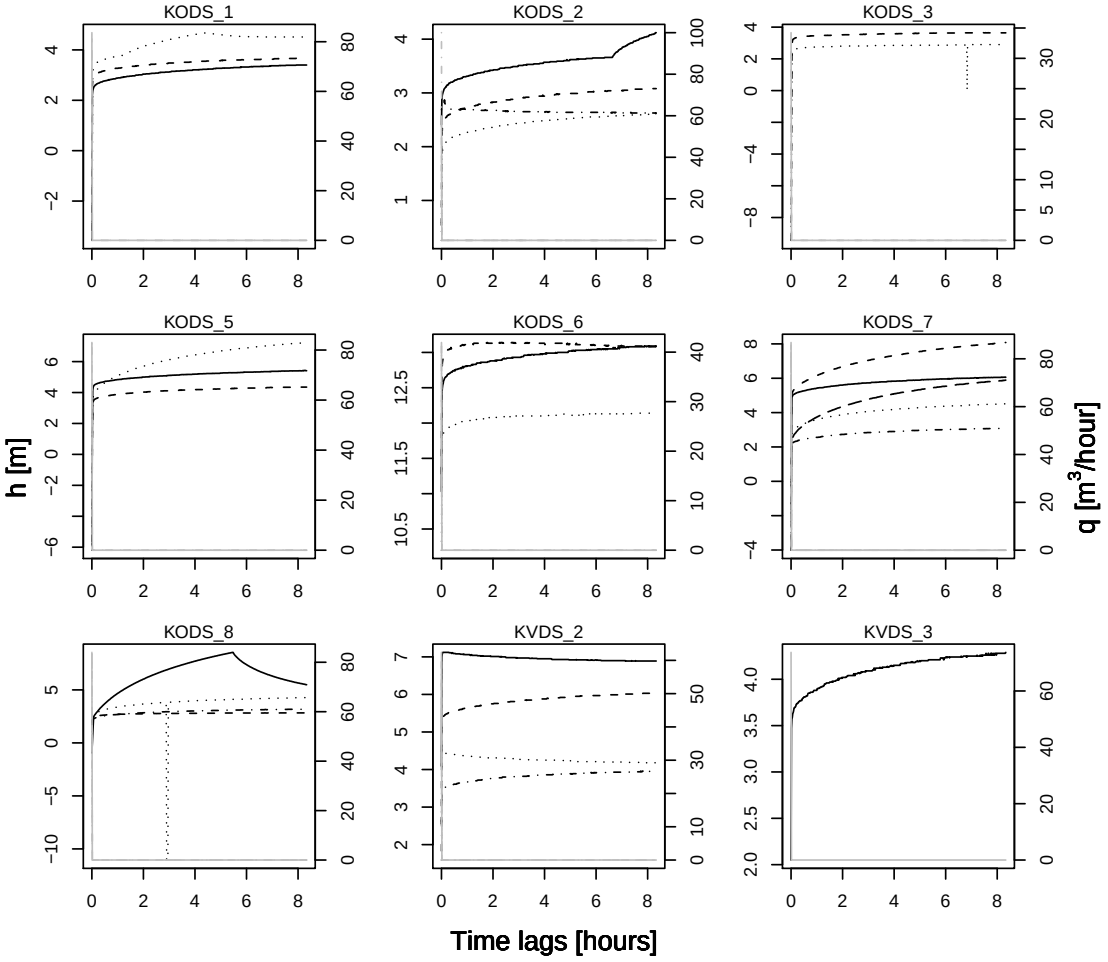


Figure 40: Step response for the individual drawdown in each well. Drawdown when corresponding pumping rate is turned off.

cates that an essential information for the drawdown is hidden in that interval. Ideally should the data be taken more frequently then one minute, especially for data points following sudden changes in discharge from some particular well. The observations are directly from the water level in the wells, hence for changes in the discharge rate the first ensuing drop is exclusively due to the discharge from the well, before the water in the aquifer adjust to the interference. However, this phenomena, occuring in first few moments after the water withdrawal from the individual well, has to be overlooked and assumptions must be based on the available data.

The above shows that there are three possible time constants in the system, although the one detected for the first minutes from decision, in both pump being switched on and off, is not direct relation to the system. However, in the grey box modelling framework consideration is taken to all three timeconstants.

Following this report is the objective to formulate the water level in the well field by only

using the available data for flow rates and water elevation in the wells. In contrast to the lack of data to describe the behavior of the watertable closest to the wells, the available data of one minute for model assessment is rather fine resolved. The variation between consecutive measurements is very small for really long periods in the data set, which implies that the data should be aggregated for the modelling procedure.

5.1 Irregular subsampling from treated time series

Subsampling the data is required to determine an adequate time series to sufficiently estimate a model structure for groundwater flow in a well field, i.e. achieve sufficient estimations for drawdowns in the well field by only considering the observed flow rates from the wells as decision variables. To reach this goal at least one unobserved state variable between any two wells has to be estimated. The waterlevels in these particular state variables are first affected by discharge in the neighboring well when water from the aquifer starts to flow to the well, i.e. sudden changes in discharge from a neigboring well has more affect on the waterlevel closest to the well and any abrupt interference regarding the well have rather limited influence on this particular state variable.

The most rapid changes happen within several minutes from the intervention of a new pumping strategy. From the data it is transparent that for such an event the following measurements for the water drawdown are crucial for all wells, since discharge from a single well is detected in the entire well field. However, in the long run it appears as the water drawdown reaches an equalibrium and the difference between consecutive water level measurements is minimal. When the drawdown reaches equalibrium point, the fluctuation in the observed values is depleted and the waterlevel is considered constant. In the observed time series this is occationally the case, but this is not sufficient for a model structure, which is merely designed to grasp the dynamics of the physical model.

Ideally, larger time steps, which consider an average over the time step, would give more feasible sampling times for the water drawdown observations when reached equilibrium. By considering the time constant, the sampling time should be close to 4 hours. On the contrary is this not suitable for the drastic influence occurring in the first subsequent minutes from a change in the decision variables.

This indicates that there is a need for irregular sampling, which depends on the decisions made for the pumps. Every time decision changes, the consecutive 20 minutes are sampled, followed by 5 minutes average sampling times. One hour after decision the data is subsampled every 10 minutes by considering the mean value over the sampling time, after 2 hours the sampling is every 30 minutes until a new decision is made. If new decision is taken within the first 2 hours the sampling rutine start from the beginning, and disregarding the remains of the previous sample schechule.

The subsampling is illustrated in Figure 41 and shows that the suggested subsampled time series adequately describes the dynamics of the water drawdown in one of the wells, for any changes in the pumping rate. The top panel shows one of the five treated time series, spanning the period of November 3rd, 2008 to March 21st, 2009, along with the

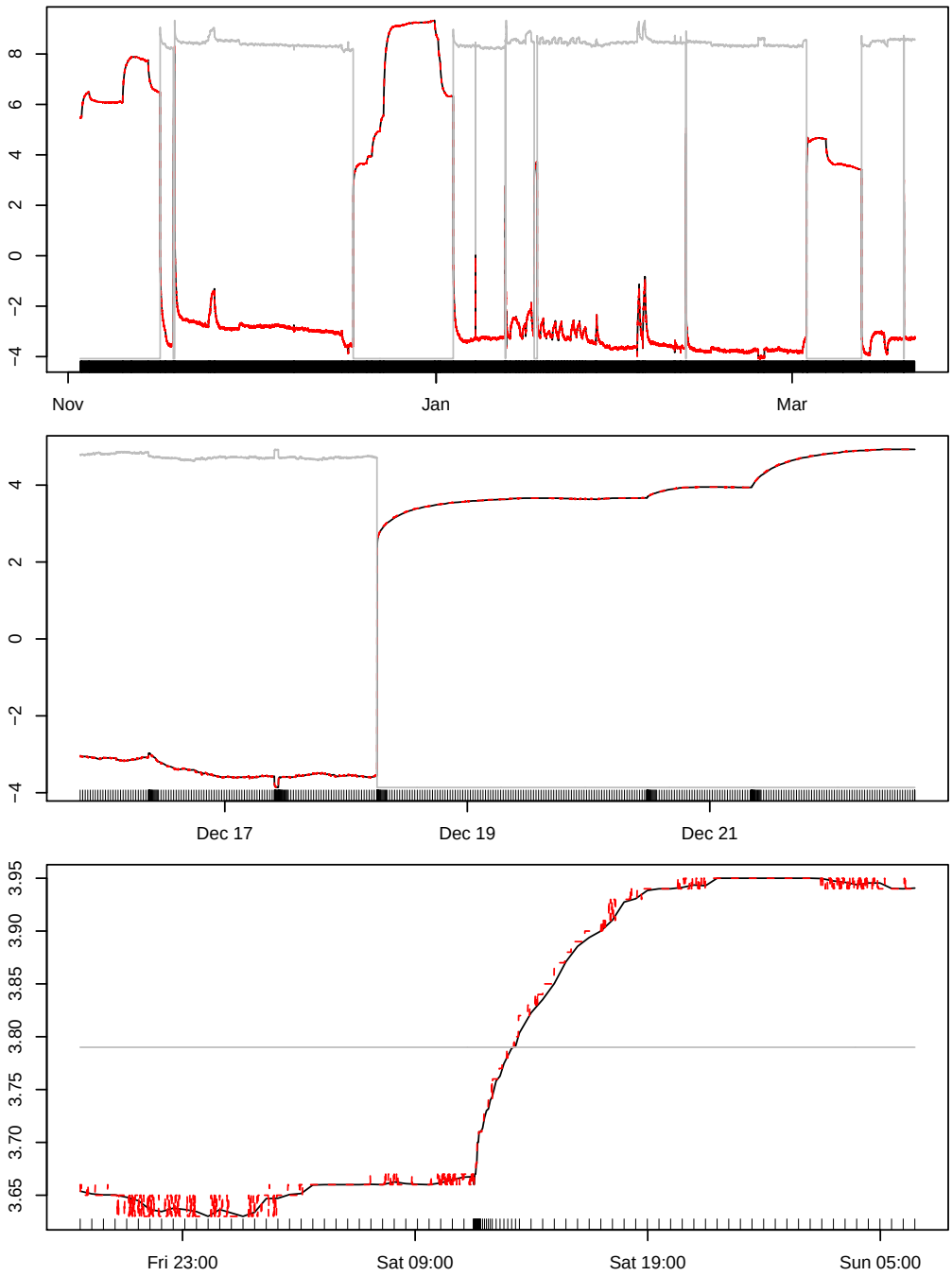


Figure 41: Comparison between the measured time series with one minute resolution (red) and the irregular subsampled time series (black) for the first time series in the KODS_1. Correpsonding pumping rates (grey) show how the drawdown is affected by it, as well as how the water level is correlated with pumping from alternative wells in the well field.

correpsonding pumping rate. It appears as the subsampled data (black solid line) is matching the observed time series. By sooming in on parts of the series (middle and bottom panels) the analogy between the two series is verified. The black tick marks at the bottom of each panel in Figure 41 shows where the data is subsampled.

For the top panel it is impossible to detect the irregular sampling time, but for the middle plot, showing sampled data for about 4 days, the subsampling time instants become a little more clear: When the pump is switched off, the water level *j*umps in the first minutes, and then continues to increase more smoothly. This is the case for changes in pumping rate for this particular well, but changes in the waterlevel related to decisions for other wells in the well field can also been seen in the plot. This relation is then verified by the bottom panel, showing changes in water drawdown regarding changes in pumping rate in some other well then the one display in the figure.

By subsampling as illustrated above, the size of the data set is reduced tremendously, though without loosing any information from the treated data sets. Number of records in the subsampled data sets are, respectively, 9,563, 1,002, 6,000, 423 and 3,966. The total number of records of the subsampled series is 20,954. Compared to the treated data series, listed in Section 3, the size of the subsampled data series is only 5% of the treated data series. The treated subsampled data sets are considered sufficient for the continuous-time stochastic model, developed for the Søndersø well field.

6 Summary and Conclusions

This report illustrates a dataset of measurements taken at Søndersø well field, the numeric treatments that have been applied on the dataset, and some preliminary data analysis. The provided dataset includes measurements for drawdown and corresponding discharge for all wells on the East side (8 wells) and West side (3 wells) of Søndersø. The total flow from the wells in the Tibberup site (South); and the total flow from the Søndersø well field. Since measurements have been taken with regular time interval over three different periods, the dataset is divided into three subset, one per period. Flaws and mismatches in the data have been detected and eliminated or treated by using interpolation or splines. Missing data have been removed from the dataset, causing the three subsets to be further divided into total of five subsets. Treated data have been re-grouped into special datasets called *Stationary Decision Datasets*, or SDD. Each SDD contains water drawdown and pumping rate time series which have been recorded under similar operational circumstances. Following the analysis of SDDs, the dynamics of pumping rate and water drawdown have been investigated and decomposed into three components; transition, oscillation and noise. In water drawdown time series, transition is always the dominating component; oscillation and noise represent the residuals. A transition is triggered when at least one pump in the system is switched on/off. The water level at each well moves, with decreasing speed, from a initial level to a new equilibrium level. The decomposition shows that all pumping rates on East and West of Søndersø, are highly correlated with the total flow from the South. The data analysis also suggests a linear representation as an adequate modelling framework for the grey box approach. This report also illustrates how time series have been subsampled. Subsampling is a fundamental step in modelling, consisting on selecting values from the dataset that will be used for calibration of the parameters in the grey-box model. The sampling technique herein adopted is based upon the fact that transitions occur in a asymptotic way, i.e. with decreasing speed. Consequently, the treated dataset is subsampled with decreasing frequency: every minute for the first 10 minutes; every 30 minutes thereafter, until the next transition. Such a subsampling approach is designed to calibrate the parameters of the grey-box model with a representative database which is also reasonably reduced in size.

References

Ahlfeld D P & Mulligan A E, 2000. *Optimal Management of Flow in Groundwater Systems*. Academic Press, San Diego, USA.

Coleman T F & Li Y, 1992. *A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables*.

Davis L D & Mitchell M, 1991. Handbook of genetic algorithms. *Van Nostrand Reinhold*.

de Boor D, 1978. *A Practical Guide to Splines*. Springer-Verlag.