Complexity versus Simplicity:

An Example of Groundwater Model Ranking with the Akaike Information Criterion

Engelhardt, I.[1), 2) *], De Aguinaga, J.G.[3)], Mikat, H.[4)],

Schüth, C[1)]., Lenz, O.[1)], Liedl, R.[5)]

1) Technische Universität Darmstadt, Institute of Applied Geosciences, Germany
2) Forschungszentrum Jülich, Institute of Bio- and Geosciences, Germany
3) Bauhaus-Universität Weimar, Research Training Group, Germany
4) Hessenwasser GmbH & Co. KG, Germany
5) Technische Universität Dresden, Institute for Groundwater Management, Germany

*Corresponding author address: i.engelhardt@fz-juelich.de

Abstract

A groundwater model characterized by a lack of field data to estimate hydraulic model parameters and boundary conditions combined with many piezometric head observations was investigated concerning model uncertainty. Different conceptual models with a stepwise increase from 0 to 30 adjustable parameters were calibrated using PEST. Residuals, sensitivities, the Akaike Information Criterion (AIC and AICc), Bayesian Information Criterion (BIC), and Kashyap's Information Criterion (KIC) were calculated for a set of seven inverse calibrated models with increasing complexity by gradually rising the number of adjustable model parameters. Finally, the likelihood of each model was computed. As expected, residuals and standard errors decreased with an increasing amount of adjustable model parameters. The model with only 15 adjusted parameters was evaluated by AIC as the best option with a likelihood of 98 %, while the model based on sedimentological information obtained the worst AIC value. BIC and KIC selected a simpler model than the model chosen by AIC as optimal. Computing of AIC, BIC, and KIC yielded the most important information to assess the model likelihood. Comparing only residuals of different conceptual models was less valuable and would result in an overparameterization and certainty loss in the conceptual model approach. Sensitivities of piezometric heads were highest for the model with five adjustable parameters following distinctively changes of extracted groundwater volumes. With increasing amount of adjustable parameters piezometric heads became less sensitive for the model calibration and remained constant during the simulated period. With increasing freedom model parameters lost their impact on the model response. Additionally, using only sedimentological data to derive hydraulic parameters possessed a consistent error into the simulation results and cannot recommended generating a true and valuable model.

Keywords: AIC, BIC, KIC, Sensitivity, Uncertainty, PEST

1. Introduction

Uncertainty is a key issue in hydrogeological modeling. Uncertainties are associated with parameter values, chosen scale, data quality, validity of boundaries, and initial conditions. Moreover, groundwater models are subject to several errors resulting from conceptual and stochastic uncertainty. Uncertainty in calibrated parameters can originate from inaccuracies in field data, insensitivity with regard to changes in model parameters, and correlations within adjusted parameter sets (Singh et al., 2010). In many cases, measured field or laboratory data cannot be directly used to parameterize the model since they are collected at different temporal or spatial scale. Overparameterized models increase uncertainty since the information of the observations is distributed through all the parameters. To simulate a natural system with a numerical model, data have to be filtered, averaged and modified. A way to reduce this uncertainty is to select a *parsimonious model*, which provides good performance with as few calibrated parameters as possible.

There are several approaches to find this compromise between model fit and low number of calibration parameters (Hill and Tiedeman, 2007, Massmann et al. 2006). One of these approaches is the Akaike Information Criterion (AIC; Akaike, 1973). AIC is a probabilistic criterion based on the maximum likelihood theory and treats the problem of parsimonious model selection as an optimization problem across a set of proposed conceptual models (Burnham and Anderson, 2002). In addition, AIC allows the ranking of models and determines the optimal model for a given data set. It identifies wherever the results of the selected model are already satisfactory or wherever an increased effort is needed by introducing more parameters into a model, so that AIC is able to select a more complicated model with a better fit to the observed data.

The application of the AIC is relatively new in groundwater modeling and still not standard, although it has been applied in several studies (e.g., Foglia et al., 2007; Hill, 2006; Hill and Tiedeman, 2007; Katumba et al., 2008; Parker et al., 2010; Poeter and Anderson, 2005; Singh et al., 2010; and Ye et al., 2010). Foglia et al. (2007) uses piezometric pressure heads and stream flow gauges for a groundwater model with a huge area of that were monitored over some month and calibrates the hydraulic conductivity. Poeter and Anderson (2005) analyzed synthetic data sets, Katumba et al. (2008) investigates the likelihood of models of tank experiments, and Parker et al. (2010) analyzes two impeller flow loggings.

Singh et al. (2010) and Ye et al. (2010) compared the model uncertainty with respect to the estimated recharge for the Yucca Mountain nuclear waste repository that is well documented over decades of years. In this study, a typical field-generated data set, as often available for numerical investigations for groundwater management issues was investigated. The data set suffers from a lack of information on boundary and initial conditions, however, observation data were collected in great quantities and over a long-term. Information criteria, such as the AIC, might be helpful to define the best model concept with respect to the model performance and uncertainty.

The investigated groundwater model was based on very few data available from pumping tests giving hydraulic properties of the aquifer and most hydraulic parameters had to be estimated from sedimentological investigations. Sedimentological information was derived from borehole drillings conducted more than 100 years ago and was associated with high uncertainties. On the other hand, long-term data, in form of high resolution groundwater level time series, were provided for the model calibration. In this study, the uncertainty of different model approaches was adressed by gradually increasing the amount of adjustable model parameters to predict measured groundwater fluctuations. Finally, the optimal model selected by different information criteria (Akaike's Information Criterion, AIC and AICc, Bayesian Information Criterion, BIC, and Kashyap's Information Criterion, KIC) were evaluated considering the calibration results and the parameter uncertainties of the model.

2. Materials and Methods

2.1 Investigated Field Site

Geological Setting

The study area is situated south of the city of Frankfurt and east of the Frankfurt International Airport in the German federal state Hesse. The site is located in the northern part of the Upper Rhine Graben (URG), which is part of the European Cenozoic Rift System (Ziegler and Dèzes, 2005). The URG, an approximately 300 km long and 40 km wide elongate lowland is flanked by uplift plateaus and terminated in the northern part by the WSW – ESE striking southern boundary fault of the Rhenish Massif, bounded to the West by the Mainz basin and to the East by the Hanau basin and the Odenwald Massif (*Fig 1a*). The graben-filling sediments are of Eocene to Early Miocene and of Plio-/Pleistocene age (Berger et al., 2005). The subsidence of the graben resulted in up to 2000 m thick Tertiary deposits and more than 100 m thick fluvial Quaternary sediments (Anderle, 1968; Bartz, 1974). In the northernmost part of the URG between Mörfelden, Langen, Frankfurt, and the Lower Main area mainly fluvial sand and gravel with embedded clay lenses were deposited during the Pleistocene (Anderle, 1968). The thicknesses of these deposits in the northern offset of the URG range between 10 and 40 m (*Fig 1b*). Holocene eolian silty fine sand was deposited on top of this layer. The base of the Quaternary and Tertiary sand and gravel consists of Permian sandstone and conglomerates as well as Tertiary basalt.

Hydrogeology and Hydrology

Average groundwater flow velocities within the Quaternary and Tertiary sand and gravel deposits are about 0.5 m/d and groundwater flows from the Sprendlinger Horst in the South-East towards the river Main. The depth to the groundwater table varies between 3 and 5 m near the river Main and gradually increases up to 15 m towards the South and East.

Fig. 1: a) Simplified geological map showing the northern part of the Upper Rhine Graben, the adjacent Mainz and Hanau basins (modified after Lahner and Toloczyki (2004); W: Wiesbaden, M: Mainz, F: Frankfurt, H: Heidelberg). b) Thickness of the Quaternary sand and gravel deposits south of Frankfurt (after Anderle, 1968; Bartz, 1974; Anderle and Golwer, 1980). Location of the model domain, the water works, and of transect A-B.

The long-term precipitation (1961-1993) averages around 675 mm/a as measured at the meteorological station in Frankfurt. About 15% of the precipitation, thus 100 to 150 mm/a, can infiltrate into the groundwater (Berthold & Hergesell, 2005). The groundwater within this area is intensively used for drinking water and industrial purposes. Several water works are located within this region. In the water works Oberforsthaus, located directly in the study area, 18 production wells were operated. Groundwater extraction started already in 1894. About 100 years later, the water works was rebuilt and then extraction rates increased within a few years from 560,000 m^3/a (1995) to $1.4*10^6$ m^3/a in 2000. Since 2005, the water works has been kept in stand-by operation. For sustainable groundwater management issues groundwater resources were recharged with treated water from the river Main to prevent an excessive groundwater table drop. Surface water was infiltrated by horizontal pipes and a small pond (named Jacobi-pond). During periods of high groundwater extraction rates treated surface water infiltration reached up to 35 to 40% of the extracted groundwater volume and was reduced to about 25% in periods with average extraction rates. The artificial groundwater recharge stopped in 2005, when the water works changed to stand-by operation.

2.2 Numerical Model Set-up

Discretization

The geological structure of the investigated Quaternary aquifer consists of a complex system of high and low permeable layers. Nine lithological units were identified in the borehole drillings. For translation of the complex geological information into a numerical model some simplifications were necessary. All geological information obtained from drillings and geological maps were summarized into three hydrostratigraphic layer (*Fig. 2*): (i) dominated by high permeable aquifer material (gravel and coarse sand), (ii) dominated by medium and low permeable aquifer material (medium and fine sand), and (iii) a deeper layer dominated again by high permeable material (gravel and coarse sand). The impermeable aquifer base is built of silt, clay, sandstone, limestone, or basalt. Then, 15 profiles were constructed containing these three hydrostratigraphic layer. Geological information between the profiles were interpolated to estimate the top and bottom of the three hydrostratigraphic layer (*Fig. 2*).

With these simplifications the spatial discretization contained 22,680 grid cells. The temporal discretization for the simulated period of 19 years, ranging between

1990 and 2009, included 379 stress periods to capture the monthly collected piezometric pressure heads.

Fig 2: Averaged hydrostratigraphic layer from nine lithologic units along transect A-B.

Hydraulic Properties

Only very few data were available about hydraulic conductivities and storage of the aquifer layers. Within a layer, several micro layers may be present and an averaging technique was applied to account for these heterogeneities. First, all data obtained from the geological description of the borehole data were used to assign an initial estimate on hydraulic conductivities and storage coefficients to each of the nine lithological units. For each of the three hydrostratigraphic layer an equivalent hydraulic conductivity and storage coefficient was calculated to account for the contribution of the lithological units within each hydrostratigraphic layer, respectively (*Fig. 3*).

Fig. 3: Averaging technique to derive the equivalent hydraulic conductivities around two wells within the three hydrostratigraphic layer that contain nine lithologic units.

As an example, the equivalent hydraulic conductivity (K_{eq}) of hydrostratigraphic layer 1 around well A was obtained by calculating the weighted arithmetic average of the lithological units with:

$$K_{eq,1} = \frac{d1A \cdot K_{gravel} + d2A \cdot K_{coarse\ sand} + d3A \cdot K_{fine\,sand} + d4A \cdot K_{gravel}}{d1A + d2A + d3A + d4A} \quad (1)$$

with $K_{eq,1}$ *= Equivalent hydraulic conductivity of layer 1*

dA = Thickness of the lithological unit in the respective hydrostratigraphic layer at well A

1,2,… = Number of the lithological unit

K = Hydraulic conductivity estimated from the sedimentological description of the lithological unit

Equivalent hydraulic conductivities and storage values were interpolated over the model domain for each of the three hydrostratigrahic layers and subdivided into ten conductivity and storage zones, respectively (*Fig. 4*). Hydraulic conductivity and storage zones showed a different pattern and frequency in each of the three layers or were not developed at all. The interpolation of the equivalent hydraulic conductivity zones failed around geological structures such as faults. Therefore, a final manual adjustment of the hydraulic parameters to maintain relevant geological features was necessary.

Fig. 4: Spatial distribution of the ten equivalent hydraulic conductivities of Model 1 (uncalibrated model based on sedimentological information) within the three hydrostratigraphic layer.

Numerical Model Boundaries

The standard finite-difference model MODFLOW (Harbaugh et al., 2005) was used for the flow simulations. Groundwater levels measured in 1990 within 47 observation wells were interpolated and used as initial head distribution (*Fig. 5*).

The main inflow into the groundwater system is recharge that varied monthly during the investigated 20 years. Further groundwater inflow was caused by surface water infiltration from the Jacobi Pond. Groundwater outflow mainly occurred by exfiltration into the river Main (*Fig. 5*). The stage of the river Main was adjusted monthly during the investigated period by applying a linear interpolation between two hydrological stations close to the model domain: Frankfurt Osthafen (4 km upstream) and Raunheim (16 km downstream). The water level of the Jacobi Pond was assumed to remain constant during the investigated period since groundwater levels measured near the pond remained fairly constant. Leakage between groundwater and surface water is driven by the gradient between the surface water stage and the groundwater, and the conductivity of the river bed and Jacobi Pond bottom sediments. The stage of the surface water was prescribed during the simulations, while the hydraulic conductivities of the river bed and Jacobi Pond sediments were adjusted in an initial manual “pre-calibration”. Along the South-West boundary, groundwater flowed out of the model domain towards the water works Goldstein, which started operation in 1995. This subsurface outflow was accounted for by a general head boundary. The piezometric head outside of the model domain was given by the monthly measured groundwater level at the pumping wells of the water works Goldstein. Within the model domain the water works Oberforsthaus operated about 18 pumping wells between 1990 and 2005. The monthly measured extraction rates were corrected by the injected artificial recharge, and resulting extraction volumes were assigned at the water works location.

Fig. 5: Boundary conditions, initial head distribution of the numerical flow model and location of the observation well groups.

Model Calibration

The non-linear parameter estimator PEST (Doherty, 2010) was used for the automated model calibration through an inverse parameter estimation process based on the Gauss-Marquardt-Levenberg method. PEST minimizes discrepancies between model simulated outputs and the corresponding measurements by minimizing the weighted sum of squared differences between the respective values. PEST also computes the sensitivities with regard to selected parameters at all observation points. These sensitivities provide a measure of how much a simulated value changes in response to a perturbation of an adjustable parameter (Hill & Tiedeman, 2007).

In PEST the composite sensitivity s_j of a parameter i is computed with (Doherty, 2010):

$$s_i = \left(\mathbf{J}^{\mathbf{t}}\mathbf{Q}\mathbf{J}\right)_{ii}^{1/2} / m \qquad (2)$$

*where **J** is the Jacobian matrix**,** **Q** is the weight matrix, $\boldsymbol{J^tQJ}$ is the normal matrix, and m is the number of observations with non-zero weights.*

The composite observation sensitivity s_j of observation j is computed in PEST with (Doherty, 2010):

$$s_j = \left\{\mathbf{Q}\left(\mathbf{J}^{\mathbf{t}}\mathbf{J}\right)\right\}_{jj}^{1/2} / n \qquad (3)$$

where $\boldsymbol{J^tJ}$ is the Hessian matrix, j is the counter of the observations, and n is the number of adjustable parameters.

Piezometric heads collected at 41 observation wells between 1990 and 2009 were used for the model calibration giving a total number of 5,081 observation points (*Fig. 5*). For a better overview, observation wells were categorized into six groups: (i) near Jacobi Pond, (ii) near the River Main, (iii) Southern area, (iv) Western area, (v) Northern area, and, (vi), around the water works Oberforsthaus (*Fig. 5*) to account for the different factors influencing the hydraulic pattern of the investigated region. Hydraulic conductivities and storage coefficients were estimated using PEST. First guesses of these parameters were assigned as derived from sedimentological interpretation of the borehole data (*Fig. 3 and Fig. 4*).

Composite observation sensitivity s_j were computed for each observation point to be an overall measure of the sensitivity of all 5,081 observation points to all adjustable parameter in the model, respectively.

After calibration of the hydraulic parameters a validation was conducted with the optimal model selected by the information criterion. This validation analyzed piezometric pressure heads measured at six further observation wells representing each observation group. These observation wells were not used in the prior parameter estimation during the inverse modeling. This procedure was chosen due to the analysis of Bredehoeft and Konikow (2012). They emphasize that a professional judgment of the model is only possible using historical data, while the validation of the model against future response remains challenging. However, errors resulting from conceptual errors will neither be addressed by using historical nor future data in the validation (Bredehoeft and Konikow, 2012).

2.3 Principles to Weight and Rank Models using AIC, AICc, BIC, and KIC

Akaike's Information Criterion

Computation of the AIC allows the selection of a parsimonious model that uses the smallest number of parameters needed to provide an adequate approximation to the measured data. Thus, a compromise between a "good" fit and a small number of parameters can be found.

Akaike (Akaike, 1973) defined a model selection criterion called Akaike's Information Criterion (AIC) that is based on the estimation of the information loss between an approximating model and an unknown parametrized truth. AIC is defined as follows (Ye et al., 2008):

$$AIC = n\ln(\hat{\sigma}_{ML}{}^{2}) + n\ln(2\pi) + n + \ln|Q^{-1}| + 2p \quad (4)$$

where p equals the number of estimated model parameters plus one, n is the number of observations, Q is the weight matrix, and $\hat{\sigma}^2_{ML}$ represents an estimate of the variance of weighted residuals, which is given by:

$$\hat{\sigma}^2_{ML} = \frac{\sum_{j=1}^{n}(\varepsilon\, q)_i^2}{n} \quad (5)$$

where ε stands for the residuals (observed minus calculated values), and q is the weight of the j^{th} observation, respectively, which is always one for the present study.

The first term in Eq. 4 represents the lack of the model fit, which decreases when more parameters are included. The last term can be seen as "penalty" term for incorporating more parameters as this term increases within rising amount of adjustable parameters.

The two middle terms are constants for a specific data set, and are not affected if parameters are added or removed from the models (Cavanaugh, 1997). Weights were set to one since no information about data uncertainty and measurement error was available. However, when additional information about confidence of the data is available the weight matrix of Eq. 4 allows comparing models based on a weighted data set of observations. This reflects the confidence to specific measurements, or simply, provides the flexibility to scale observations according to additional information or normalization procedures (Hill and Tiedeman, 2002).

Akaike (1978) defined weights w_j to obtain a relative measure of the likelihood of a model for a given set of N models. These weights are expressed as:

$$w_j = \exp(-0.5\Delta_j) / \sum_{j=1}^{N} \exp\left(-0.5\Delta_j\right) \qquad (6)$$

where j is the counter of models, and $\Delta j = AIC_j - AIC_{min}$ *denotes the AIC difference to the smallest AIC of all considered models.*

The larger the AIC difference of a model, the less likely it is to be the best one.

Alternative Information Criteria

Several modifications of AIC have been developed. For the case of having a small sample, n/K<40, Burnham and Anderson (2002) suggest using AIC_c:

$$AIC_c = AIC + \frac{2K(K+1)}{n-K-1} \qquad (7)$$

where AIC is the Akaike Information Criterion as defined by Eq. 4, and K is the number of estimable parameters.

AIC_c tends to AIC when the number of observations is high relative to the number of calibrated parameters as in our study, where n/K equals 5,081/30 giving 169.

Further information criteria were also computed to provide a contrast analysis to the results obtained by the AIC. The BIC (Bayesian Information Criterion) gives a response to the concern that AIC sometimes promotes the use of more parameters than required (Hill and Tidemann, 2007). The BIC is calculated with (Doherty, 2012):

$$BIC = n\ \ln\left(\hat{\sigma}^2\right) + p\ln(n) \tag{8}$$

The KIC (Kashyap's Information Criterion) additionally considers the likelihood of the parameter estimates in light of their prior values and contains a Fisher information matrix term that imbues it with model selection properties not used by AIC, AICc or BIC. KIC weights and ranks alternative models with respect to the models' predictive performance under cross validation with real hydrologic data (Ye et al., 2008). KIC was derived in the Bayesian context by Kashyap (1982) and is calculated with (Doherty, 2012):

$$KIC = \left(n - (p-1)\right)\ln\left(\hat{\sigma}^2\right) - (k-1)\ln(2\pi) + \ln\left|\mathbf{J}^{\mathbf{t}}\mathbf{QJ}\right| \tag{9}$$

Conceptual Approach

All models were calibrated to the same data set of piezometric pressure heads, and the model with the smallest information criterion is regarded as the optimal one of all proposed models as selected by AIC, AICc, BIC, and KIC, respectively.

First, the uncalibrated model using only sedimentological information was simulated (Model 1), then the five most widespread horizontal hydraulic conductivities were estimated (Model 2). In Model 3, all horizontal hydraulic conductivities were considered and vertical hydraulic conductivities were tied by a factor of 0.1 ($K_v = K_H/10$). The next model (Model 4) computed additionally to the horizontal hydraulic conductivity the five most widespread storage coefficients. Model 5 estimated all horizontal conductivities and storage coefficients. In Model 4 and 5 vertical hydraulic conductivities were still tied. Then in Model 6 all horizontal and vertical conductivities were estimated independently and in addition the five most widespread storage coefficients. Finally, Model 7 independently estimated all horizontal and vertical hydraulic conductivities and all storage coefficients for all zones of the model domain giving a total amount of 30 adjustable parameters (Tab. 1).

Tab 1: Calibrated models analyzed with AIC, AICc, BIC, KIC.

Finally, using the paired model methodology (Doherty and Christensen, 2012) the benefit of a more complex model associated with good calibration results versus a simple model yielding a higher certainty is assessed. Simulation results of both models are given against each other in a scatter plot. Coefficients (intercept and slope) of the regression line allow analyzing the bias of the simple versus the results obtained by the optimal and more complex model with a higher degree of freedom and uncertainty.

3. Results

3.1 Sensitivity Analysis

For each observation group time-dependent dimensionless sensitivity coefficients of the measured piezometric pressure heads are shown in Fig. 6. The relative pattern of the sensitivities between the groups is independent from the number of parameters used in the automated model calibration. Sensitivity is always highest for the Northern area as the best optimization results could be obtained for this region. Lowest sensitivities are always computed for observation wells near the River Main and the Jacobi Pond (*Fig. 6*). These low sensitivities display the impact of surface water-groundwater exchange on the groundwater level. in this In this part of the study area, groundwater levels were mostly driven by the stage of the surface water and leakage through the colmation layer of the river and pond bed sediments and that were not adjusted in the automated model calibration. Hydraulic conductivities of the colmation layers of the Jacobi-pond and Main were derived from a manual “pre-calibration” and fixed with $5 \cdot 10^{-6}$ m s^{-1} and $1.2 \cdot 10^{-5}$ m s^{-1}, respectively.

Sensitivities of all observation groups follow changes of the groundwater level fluctuations and decrease, when the groundwater extraction stopped in 2005.

Fig. 6: Sensitivity of the six observation groups with respect to the adjustable amount of parameters and the cumulative groundwater extraction at the water works Oberforsthaus.

Sensitivities were compared for the seven models differing by the number of adjustable parameters from initially 5 to finally 30 parameters. The PEST optimization with five adjustable parameters revealed highest sensitivity coefficients (*Fig. 5*). Increasing the number of adjustable parameters decreased the sensitivity of the piezometric heads. Therefore, considering a model set-up with large numbers of observation data, the number of adjusted model parameters must be chosen parsimonious to prevent an overparameterization and to maintain the influence of the measured data on the model response.

3.2 Comparative Results of the Model Selection Criteria

Four information criteria were computed to select the optimal model approach. Computing the AIC, AICc, BIC, KIC allowed the evaluation of the best conceptual model with respect to complexity and parameter uncertainty. Since Eq. (4) has to be minimized, the lowest information criterion value indicates the best model. Model complexity was gradually increased from the uncalibrated stage (based on sedimentological information) to 30 adjustable model parameters (Tab. 1). This increase in complexity was linearly penalized; as expected, by considering more parameters the model fit steadily improved until reaching a constant level with only little further improvement (Fig. 7). By comparing the model fit and the penalty against the value of the information criterion the models can be ranked. The scale of the y-axis is omitted in Fig. 7 since the information criterion is a relative measure and the absolute values are meaningless. Important are, however, the differences to the best model (AIC Δ_j ; BIC Δ_j *;* KIC Δ_j *; Tab* 2)

Both, AIC and AICc assess the similar model as optimal. The lowest AIC and AICc value is achieved by Model 4 with 15 adjustable parameters. The selection of AIC and AICc mirrors the trend of the model fit that improved distinctively between Model 1 and Model 4, and stagnated with more than 15 adjustable model parameters. Model 2 (5 adjustable parameters) and Model 7 (30 adjustable parameters) were assessed similarly poor due to a lack of model fit to the data (Model 2) or an unjustified complexity (Model 7).

Fig. 7: AIC (diamond), AICc (square), BIC (triangle), KIC (circle) assessment of the calibrated models with respect to complexity and model fit.

Relative Akaike weights (AIC w_j), Eq. (4), were computed for all models to express in percent the likelihood of a model, where a likelihood of 100 % means that the corresponding model alone is regarded to represent the "best option", while a likelihood of 0 % corresponds to a model that has absolutely no support when compared to other models (*Tab. 2*). In our case, the model selected as optimal (AIC Δ_j = 0) is associated with a likelihood of about 98 %. All other models have practically no support according to the AIC and are either underparameterized (Models 1, 2 and 3) or clearly overparametrized (Models 5, 6 and 7).

Tab 2: Differences Δ_j of the AIC, BIC and KIC values to the optimal model, respectively, and likelihood of the flow models from the Akaike weights (AIC w_j).

All information criteria (AIC, AICc, BIC, and KIC) selected Model 1 (uncalibrated model based on sedimentological information) as worst model (highest information criteria). However, differences occurred in the selection of the optimal model and model ranking (Fig. 7).

The BIC assesses a very simple model, Model 2 with 5 adjustable parameters, as the optimal model and Model 7 (30 adjustable parameters) as unfeasible. BIC values of the different models are varying more pronounced than AIC values differ (Tab. 2). The KIC evaluates Model 3 (10 adjustable parameters) as optimal model and also Model 7 (30 adjustable parameters) as worst model. BIC and KIC choose as best model approaches with fewer adjustable parameters as they assume that in the true model still the prior information exist (Burnham and Anderson, 2004). Thus, they select for greater certainty, which threatens to capture a precise, but less accurate answer than that selected by AIC. Also due to a decreasing sensitivity of the observation data with increasing parameter freedom, Model 3, as selected by KIC, might still provide a valuable model concept with a reasonable precise match of the observation data. Finally, all selection criteria argue against increasing the model complexity to more than 15 adjustable parameters.

The model based solely on sedimentological information is assessed by all information criteria as worst model. The bias and worth of this simple model can be explored in detail with the paired model methodology as given in Doherty and Christensen (2012). The model output of the simple uncalibrated model is compared against the results of the optimal model selected by the AIC (Fig. 8). The regression coefficients (intercept and slope) of the line through the scatter plot allow addressing effects of simplification on the model predictions. The intercept differs distinctively from zero indicating the null space contribution of the parameter matrix to the prediction error and thus that the simple model possess consistent an error into the predictions (Doherty and Christensen, 2012). The slope of the scatter line is near 1. Hence, parameter surrogacy does not affect the uncalibrated model's ability to predict the piezometric pressure heads. The correlation coefficient of 0.99 indicated that the model based on sedimentological information might give already reasonable results. However, due its null space contribution to the prediction error the uncalibrated model based on sedimentological information can be excluded to provide already a true model.

Fig 8: Paired model analysis: predicted piezometric pressure heads of Model 1 (based on sedimentological information) versus the results of the optimal model selected by AIC (Model 4), regression line equation, and correlation coefficient (R^2).

3.3 Optimization Results

Obtained residuals

The model calibration was based on piezometric heads measured monthly between 1990 and 2009 in 41 observation wells. Measurements were not available every month at every observation well, giving a total amount of 5,081 piezometric pressure head data for the calibration. Computed and measured piezometric heads of the model with the smallest AIC (Model 4) are compared for each observation group in *Fig. 9*.

Within the Western area groundwater levels varied over 3 m. This fluctuation resulted from the impact of the water works Goldstein located south of this region. Within the southern part groundwater levels varied up to 2.1 m and also displayed the impact of the water works Goldstein. Around the water works Oberforsthaus groundwater levels varied over a range of 1.2 m. This lower groundwater level drop can be explained by the artificial recharge infiltrated at this water works. Within the Northern area, near the River Main and the Jacobi pond, groundwater levels remained almost constant with only minor fluctuations associated with changes in precipitation and river discharge during the year.

Fig 9: Simulated piezometric heads of Model 4 (optimal model) versus measured piezometric heads between 1990 and 2009. Observation wells were summarized in six groups. One observation well of each group is illustrated within the figure.

Within most regions, measured groundwater levels were reasonably well reproduced by the flow model (*Tab. 3, Fig. 9*). The smallest standard error of the weighted residuals was obtained with 0.22 to 0.23 near the Jacobi pond (group 5; *Tab. 3*). Around the water works Oberforsthaus (group 1) and within the Western part (group 4) computed standard errors of the weighted residuals increased to 0.47 to 0.51, which can still be assessed as sufficient with respect to the high uncertainties in boundary conditions and model parameter values. Calibration results obtained for observation wells located near the river Main (group 6) showed the highest standard error of the residuals with up to 1.34 that might result from the interpolation of the river stage within the model domain.

Six observation wells were used for the model validation containing 1,445 observations or 22% of initial available calibration data. Groundwater levels simulated by the optimal model matched measured values at most locations reasonably well (group 1, 4, 5, and 6) and demonstrated that model parameters were estimated within a reliable range (Tab. 3). However, at two locations (group 2 and 3) the model fit was distinctively poor and with a similar standard error as obtained with the model based on sedimentological information.

In summary, by increasing the amount of adjustable hydraulic conductivities, mean residuals decreased and the standard error of weighted residuals improved from 1.18 (Model using only sedimentological information) to finally 0.74 in Model 7 with 30 adjustable model parameters.

Tab 3: Standard error of the weighted residuals of the six observation groups and total sum of squared weighted residuals for each of the seven conceptual models obtained by the inverse PEST model (Model 1 to 7) and with the AIC optimal model (Model 4) during the model validation.

Obtained Parameter Estimates

Very limited information was available from field investigations about hydraulic conductivity and storage. The ratio between vertical and horizontal hydraulic conductivities was assumed to be 1:10 within Models 2, 3, 4. Applying this assumption and additionally calibrating the most widespread storage coefficients (Model 4) was assessed by the AIC as the most certain model with a likelihood of 98 %. Hydraulic conductivities were estimated distinctively higher by PEST in most regions than derived from sedimentological information (*Tab. 4*). These differences may result from the impact of secondary flow pathways or local heterogeneities that were missed by the interpretation of the borehole data.

Tab. 4: Comparison of the initial guesses of the hydraulic conductivity based on sedimentological information and values estimated by PEST for the AIC optimal model (Model 4)

4. Concluding Remarks

The investigated combination of model parameters and calibration data could lead to an overparameterized conceptual model. A sensitivity analysis clearly demonstrated that the sensitivity at all observation points decreased by increasing the number of adjustable parameters. This reduced the influence of the field data as constrain for the model predictions. Computing the AIC, AICc, BIC, and KIC allowed the evaluation of the benefit adjusting high numbers of model parameters. The simplest model based on sedimentological information as well as the complex models were rejected by all information criteria since they are likely to be under- or overparameterized. The paired model methodology also displays the high bias possessed by the simple model into the model predictions.

Differences prevail in the choice of the optimal model. AIC selects as best model a model of “medium complexity”. It adjusted five of ten storage coefficients and all ten horizontal conductivities, while keeping the vertical conductivities tied by one order of magnitude lower. The results of the optimal model selected by the AIC approximately resemble observed hydraulic piezometric heads, while keeping estimated model parameters at a minimum. The AIC was able to maintain parsimony and makes predictions with a reasonable uncertainty. KIC and BIC give preference to simpler models increasing the model certainty and to maintain prior information. The optimal models selected by BIC and KIC adjusted only five or ten hydraulic conductivities, respectively, while storage coefficients are kept as deduced from the sedimentological investigations. The model fit is unacceptable in the optimal model selected by BIC. The KIC might be able to select the optimal model for an aquifer system that is described by more precise and well-known field data about model parameter than they were available at our study site. However, in situations with poor information about model parameter and boundary conditions the AIC selection should be given preference as it chooses a parsimony model, but with a sufficient freedom to receive an acceptable model fit. The choice made by AIC reflects the data available for calibration better than the optimal models chosen by the KIC and BIC. In our case, where extensive observation data were available, computing the AIC, and eventually the KIC, can improve model confidence, as it avoids an under- or overparameterization of conceptual models for a given data set. However, to decide between the optimal model selected by the AIC and KIC, respectively, the modeler still needs an overview about the data types converted to boundary and initial conditions and model parameters, which is disregarded in the model ranking by all information criteria.

Acknowledgements

We thank Elke Duhr and Stefan Pohl (both Hessenwasser GmbH & Co KG) and Sebastián Fernández for preparing and compiling the geological and hydraulic data. Thanks are due to the comments and suggestions of the two anonymous reviewers that helped to improve the manuscript.

5. References

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle In: Kotz S, Johnson NL, editors. (1992) Breakthroughs in Statistics, vol. 1: Foundations and basic theory. New York, USA: Springer-Verlag. pp. 610-624 ISBN: 0387975-667.

Anderle, H.-J. 1968. Die Mächtigkeiten der sandig-kiesigen Sedimente des Quartärs im nördlichen Oberrhein-Graben und der östlichen Untermain-Ebene.- Notizbl. hess. L.-Amt Bodenforsch., 88: 185-196, Wiesbaden.

Anderle, H.-J., Golwer, A. 1980. Tektonik. In: Erläuterungen Geol. Kt. Hessen 1:25000, Bl. 5917 Kelsterbach, S.50-64, 4 Abb.; Wiesbaden.

Bartz, J. 1974. Die Mächtigkeit des Quartärs im Oberrheingraben.- In: Illies, J.H. & Fuchs, K. (eds.): Approaches to Taphrogenesis: 78 - 87, Stuttgart (Schweitzerbarth).

Berger, J.-P., Reichenbacher, B., Becker, D., Grimm, M., Grimm, K., Picot, L., Storni, A., Pirkenseer, C., Derer, C., and Schaefer, A. 2007. Paleogeography of the Upper Rhine Graben (URG) and the Swiss Molasse Basin (SMB) from Eocene to Pliocene: International Journal of Earth Sciences, 94: 697-710.

Berthold, G., Hergesell, M. 2005. Flächendifferenzierte Untersuchungen zu möglichen Auswirkungen einer Klimaänderung auf die Grundwasserneubildung in Hessen. INKLIM 2012 - Integriertes Klimaschutzprogramm. Abschlussbericht Hessisches Landesamt für Umwelt und Geologie, Wiesbaden.

Bredehoeft, J. D. and Konikow, L. F. 2012. Ground-Water Models: Validate or Invalidate. Ground Water, 50 (4): 493–495.

Burnham, K.P., Anderson, D.R. 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd ed. Springer-Verlag. ISBN 0-387-95364-7.

Burnham, K.P., Anderson, D.R. 2004. Multi-model inference: Understanding AIC and BIC model selection. Sociological Methods and Research. 33 (2): 261–304.

Cavanaugh, J. E. 1997. Unifying the derivations of the Akaike and corrected Akaike information criteria. Statistics & Probability Letters, 33: 201-208.

Doherty, J. 2012. Addendum to the PEST manual. Watermark Numerical Computing: Brisbane, Australia.

Doherty, J. 2010. PEST - Model-Independent Parameter Estimation. User's Manual, 5th ed.; Watermark Numerical Computing: Brisbane, Australia.

Doherty, J., Christensen, S. 2012: Use of paired simple and complex models to reduce predictive bias and quantify uncertainty. Water Resources Research. 47, W12534.

Foglia, L., Mehl, S.W., Hill, M.C., Perona, P., Burlando, P. 2007. Testing Alternative Ground Water Models Using Cross-Validation and Other Methods, Ground Water, 45(5): 627-641.

Harbaugh, A.W. 2005. MODFLOW-2005, the U.S. Geological Survey modular groundwater model - the Ground-Water Flow Process: U.S. Geological Survey Techniques and Methods 6-A16, US.

Hill, M. C. 2006. The practical use of simplicity in developing ground water models. Ground Water, 44(6):775-81.

Hill, M. C., Tiedeman, C. R. 2002. Weighting observations in the context of calibrating groundwater models. IAHS-AISH Publication; 277: 196-203.

Hill, M.C., Tiedeman, C.R. 2007. Effective groundwater model calibration - with analysis of data, sensitivities, predictions, and uncertainty. Hoboken, USA: John Wiley & Sons. ISBN: 978-0-471-77636-9.

Kashyap, R. L. 1982. Optimal choice of AR and MA parts in autoregressive moving average models, IEEE Trans. Pattern Anal. Machine Intell., 4(2), 99–104.

Kazumba, S., Gideon, O., Yusuke, H., Kohji, K. 2008. Lumped model for regional groundwater flow analysis, Journal of Hydrology, 359 (1–2): 131-140. Lahner, L., Toloczyki, M. 2004. Geowissenschaftliche Karte der Bundesrepublik Deutschland 1: 2.000.000, Geologie, Bundesanstalt für Geowissenschaften und Rohstoffe; Hannover.

Massmann, C., Birk, S.; Liedl, R; Geyer, T. 2006. Identification of hydrogeological models: application to tracer test analysis in a karst aquifer. In: Calibration and Reliability in Groundwater Modelling: From Uncertainty to Decision Making (Proceedings of ModelCARE’2005, The Hague, The Netherlands, June 2005) IAHS Publ. 304, 2006.

Parker, A. H., West, L. J., Odling, N. E. and Bown, R. T. 2010. A Forward Modeling Approach for Interpreting Impeller Flow Logs. Ground Water, 48: 79–91.

Poeter, E.P., Anderson, D.R. 2005. Multimodel ranking and inference in ground water modeling. Ground Water, 43(4):597-605.

Singh, A., Mishra, S., Ruskauff, R. 2010. Model averaging techniques for quantifying conceptual model uncertainty. Ground Water, 48 (5): 701-715.

Ye, M., Meyer, P.D., Neuman, S.P. 2008. On model selection criteria in multimodel analysis. Water Resources Research, 44(3): W03428.

Ye, M., Pohlmann, K. F., Chapman, J. B., Pohll, G. M. and Reeves, D. M. 2010. A Model-Averaging Method for Assessing Groundwater Conceptual Model Uncertainty. Ground Water, 48: 716–728.

Ziegler, P.A. and Dèzes, P. 2007. Evolution of the lithosphere in the area of the Rhine Rift System: International Journal of Earth Sciences, 94: 594-614.

Tab 1: Calibrated models analyzed with AIC, AICc, BIC, KIC.

Model	**Number of adjusted parameters during automated model calibration**	
	Conductivities	Storage coefficients
1	based on sedimentological data	
2	5	0
3	10	0
4	10	5
5	10	10
6	20	5
7	20	10

Tab 2: Differences Δ_j of the AIC, BIC and KIC values to the optimal model respectively, and likelihood of the flow models from the Akaike weights (AIC w_j).

Model	AIC Δ_j	BIC Δ_j	KIC Δ_j	AIC w_j
1	4698	4823	4645	0.00
2	37.8	0.0	21.8	0.00
3	10.1	4.9	0.0	0.01
4	0.0	27.5	2.8	0.98
5	8.8	68.9	20.8	0.01
6	18.9	111.8	19.7	0.00
7	29.8	155.3	50.8	0.00

Tab 3: Standard error of the weighted residuals of the six observation groups and total sum of squared weighted residuals for each of the seven conceptual models obtained by the inverse PEST model (Model 1 to 7) and with the AIC optimal model (Model 4) during the model validation.

Model	**Standard error of weighted residuals [-]**						
	G1	*G2*	*G3*	*G4*	*G5*	*G6*	*Total Residual*
1	0.734	1.966	0.752	0.728	0.265	1.393	1.18
2	0.472	0.640	0.499	0.494	0.219	1.341	0.750
3	0.470	0.607	0.628	0.503	0.233	1.306	0.745
4	0.475	0.602	0.591	0.507	0.225	1.317	0.745
5	0.473	0.599	0.636	0.509	0.234	1.307	0.744
6	0.473	0.600	0.628	0.508	0.234	1.305	0.744
7	0.472	0.613	0.574	0.508	0.217	1.306	0.743
Validation	0.397	0.844	0.990	0.509	0.317	1.092	-

Group 1: Around water works *Oberforsthaus*
Group 2: Southern area
Group 3: Northern area
Group 4: Western area
Group 5: Near Jacobi Pond
Group 6: Near river Main
Total residuals: obtained for 5,081 piezometric pressure head data
Validation: Residuals obtained for 1,445 piezometric pressure head data with the optimal model (Model 4)

Tab 4: Comparison of the initial guesses of the hydraulic conductivity based on sedimentological information and values estimated by PEST for the AIC optimal model (Model 4)

Zone	Hydraulic conductivity [m/s]			
	Horizontal		Vertical	
	Estimated from sedimentological information	*Estimated by PEST*	*Estimated from sedimentological information*	*Estimated by PEST*
1	$5.6\cdot10^{-3}$	$1.7\cdot10^{-1}$	$5.6\cdot10^{-4}$	$1.7\cdot10^{-2}$
2	$3.8\cdot10^{-3}$	$4.8\cdot10^{-1}$	$3.8\cdot10^{-4}$	$4.8\cdot10^{-2}$
3	$5.3\cdot10^{-3}$	$1.5\cdot10^{-1}$	$5.3\cdot10^{-4}$	$1.5\cdot10^{-2}$
4	$6.8\cdot10^{-3}$	$3.5\cdot10^{-2}$	$6.8\cdot10^{-4}$	$3.5\cdot10^{-3}$
5	$8.3\cdot10^{-3}$	$5.7\cdot10^{-3}$	$8.3\cdot10^{-4}$	$5.7\cdot10^{-4}$
6	$9.8\cdot10^{-3}$	$1.8\cdot10^{-2}$	$9.8\cdot10^{-4}$	$1.8\cdot10^{-3}$
7	$1.1\cdot10^{-2}$	$2.0\cdot10^{-2}$	$1.1\cdot10^{-3}$	$2.0\cdot10^{-3}$
8	$1.3\cdot10^{-2}$	$6.8\cdot10^{-2}$	$1.3\cdot10^{-3}$	$6.8\cdot10^{-3}$
9	$1.4\cdot10^{-2}$	$6.6\cdot10^{-2}$	$1.4\cdot10^{-3}$	$6.6\cdot10^{-3}$
10	$1.0\cdot10^{-7}$	$4.3\cdot10^{-7}$	$1.0\cdot10^{-8}$	$4.3\cdot10^{-8}$

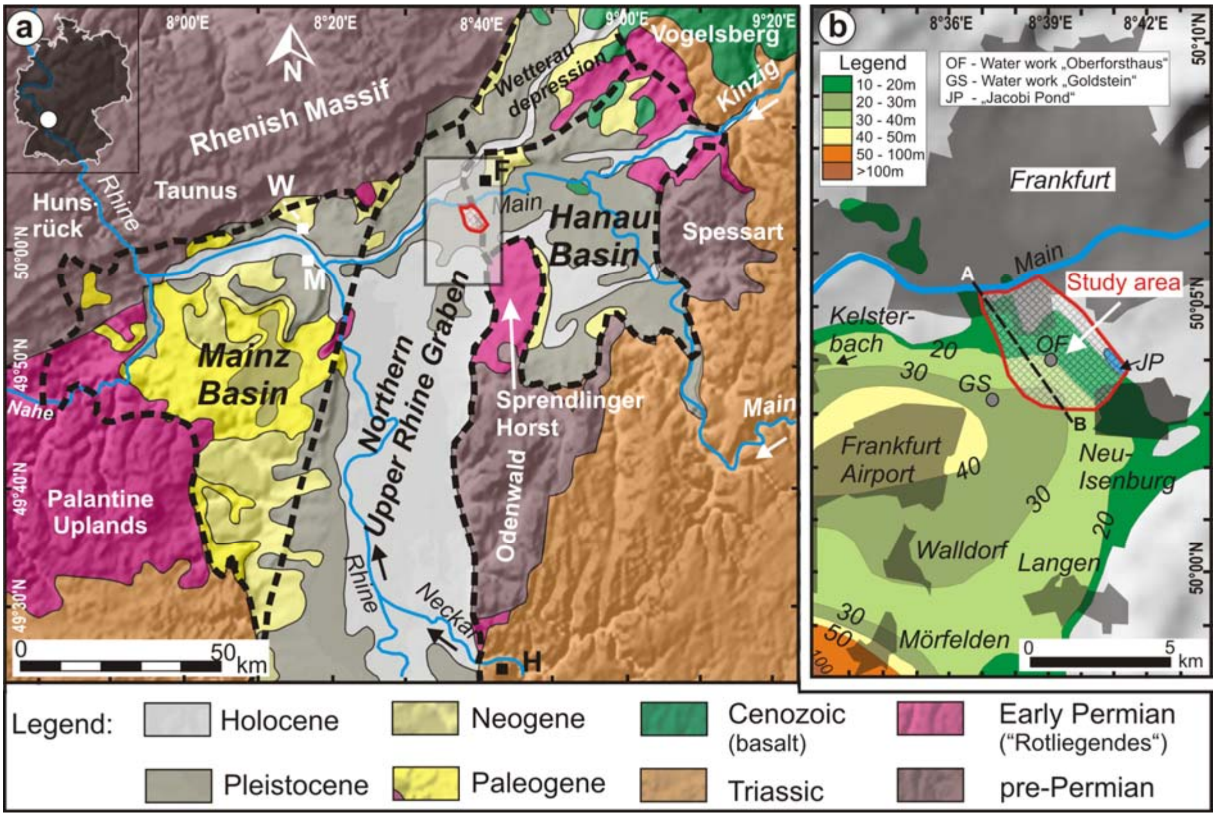

Fig. 1: a) Simplified geological map showing the northern part of the Upper Rhine Graben, the adjacent Mainz and Hanau basins (modified after Lahner and Toloczyki (2004); W: Wiesbaden, M: Mainz, F: Frankfurt, H: Heidelberg). b) Thickness of the Quaternary sand and gravel deposits south of Frankfurt (after Anderle, 1968; Bartz, 1974; Anderle and Golwer, 1980). Location of the model domain, the water works, and of transect A-B.

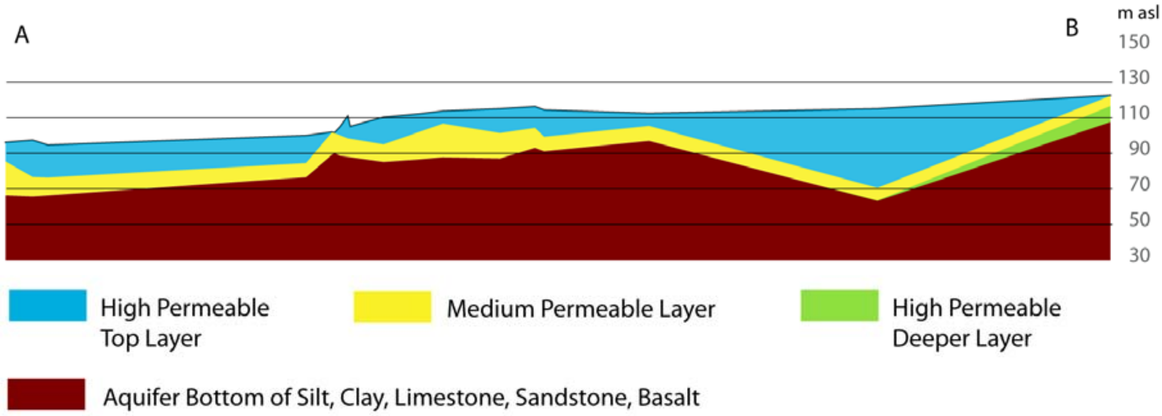

Fig 2: Averaged hydrostratigraphic layer from nine lithologic units along transect A-B.

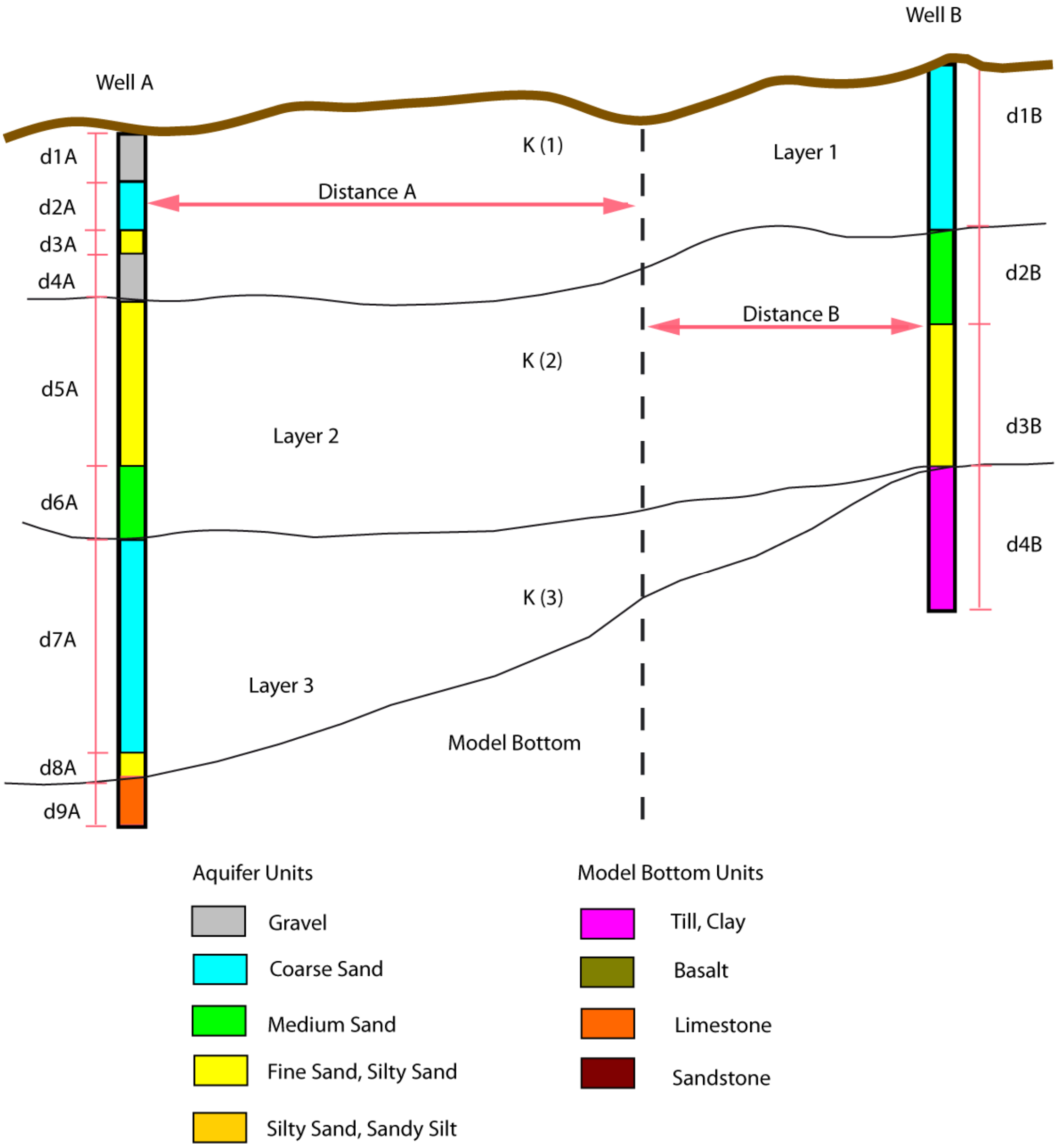

Fig 3: Averaging technique to derive the equivalent hydraulic conductivities around two wells within the three hydrostratigraphic layer that contain nine lithologic units.

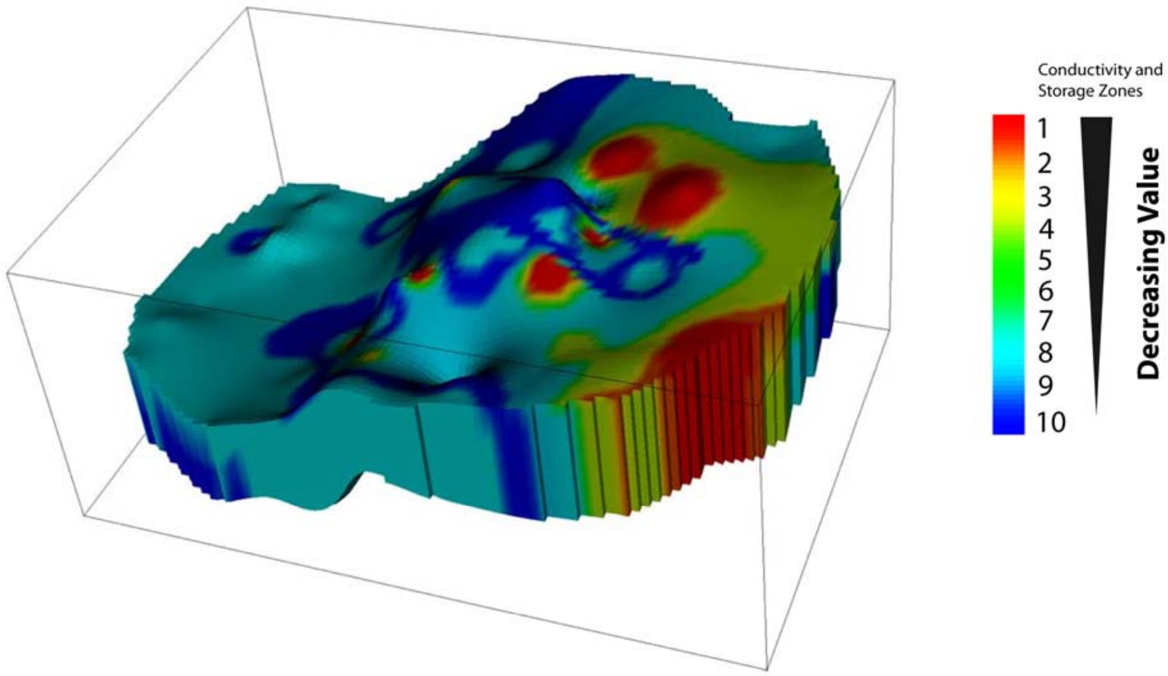

Fig 4: Spatial distribution of the ten equivalent hydraulic conductivities of Model 1 (uncalibrated model based on sedimentological information) within the three hydrostratigraphic layer.

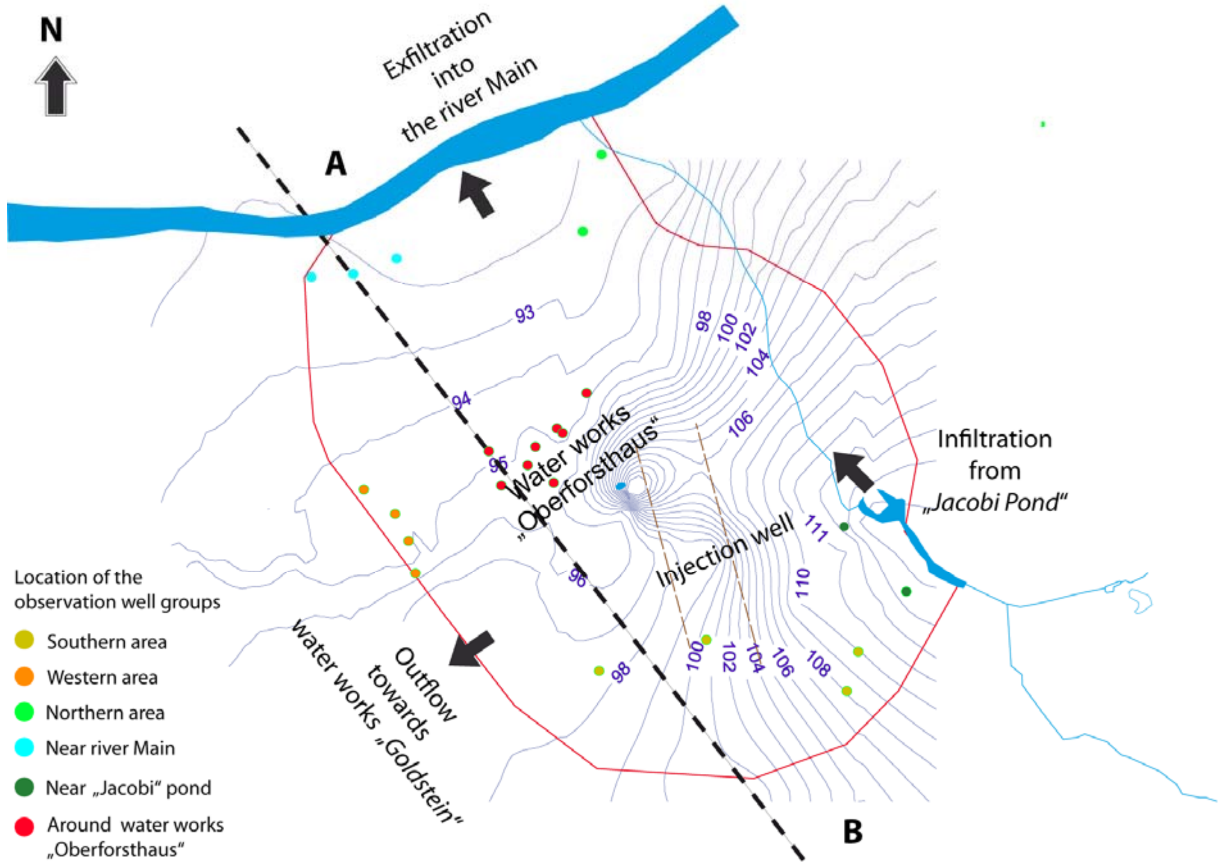

Fig 5: Boundary conditions, initial head distribution of the numerical flow model and location of the observation well groups.

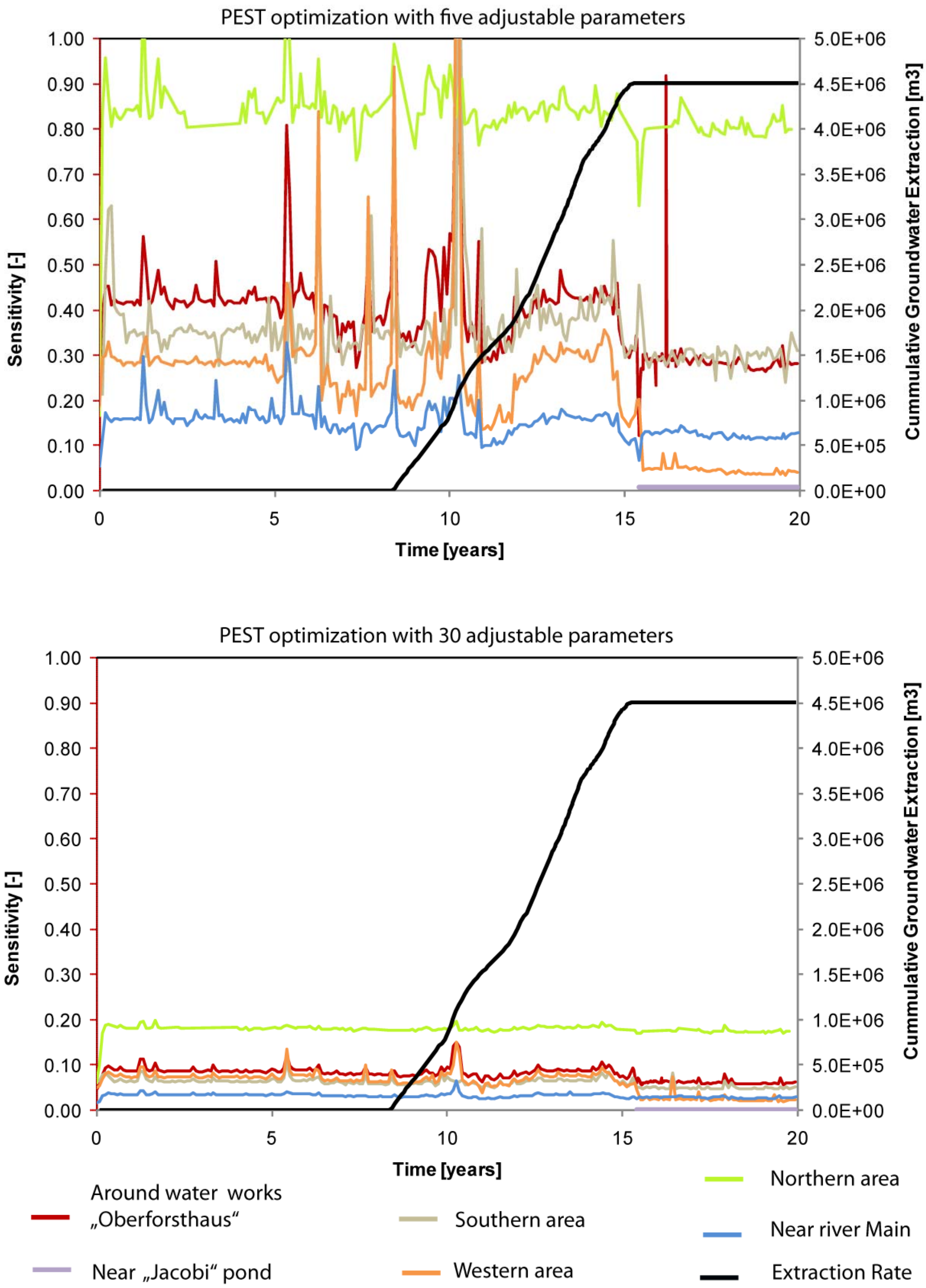

Fig 6: Sensitivity of the six observation groups with respect to the adjustable amount of parameters and the cumulative groundwater extraction at the water works Oberforsthaus.

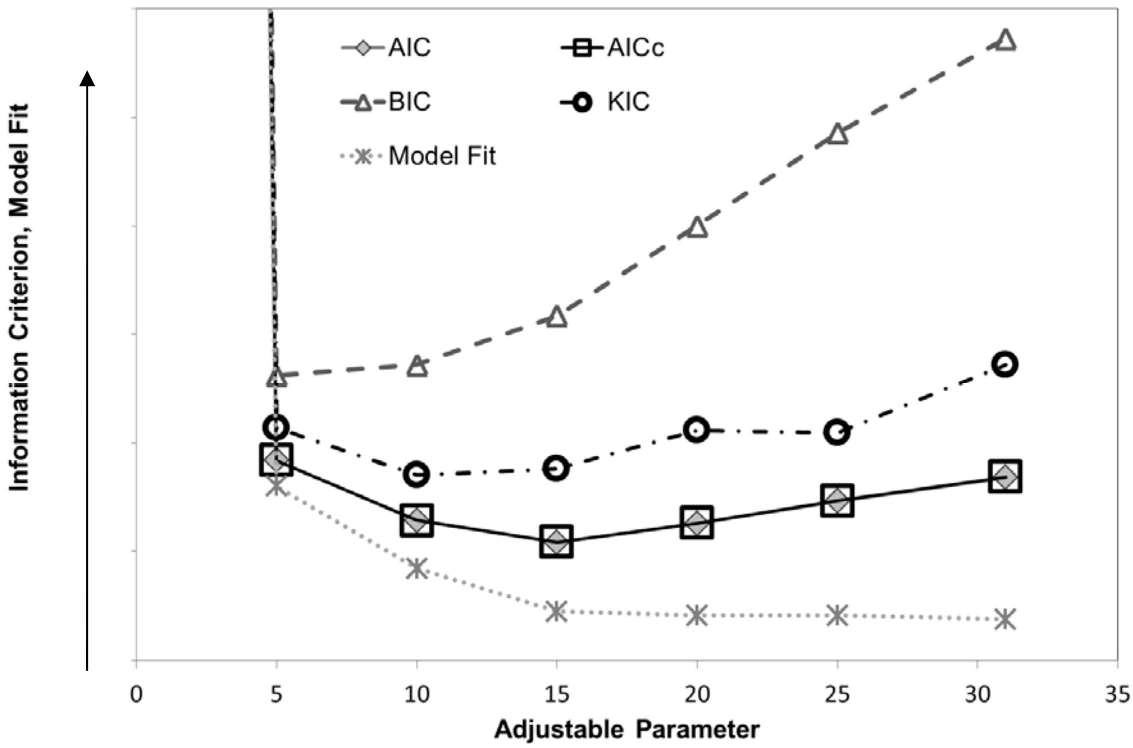

Fig 7: AIC (diamond), AICc (square), BIC (triangle), KIC (circle) assessment of the calibrated models with respect to complexity and model fit.

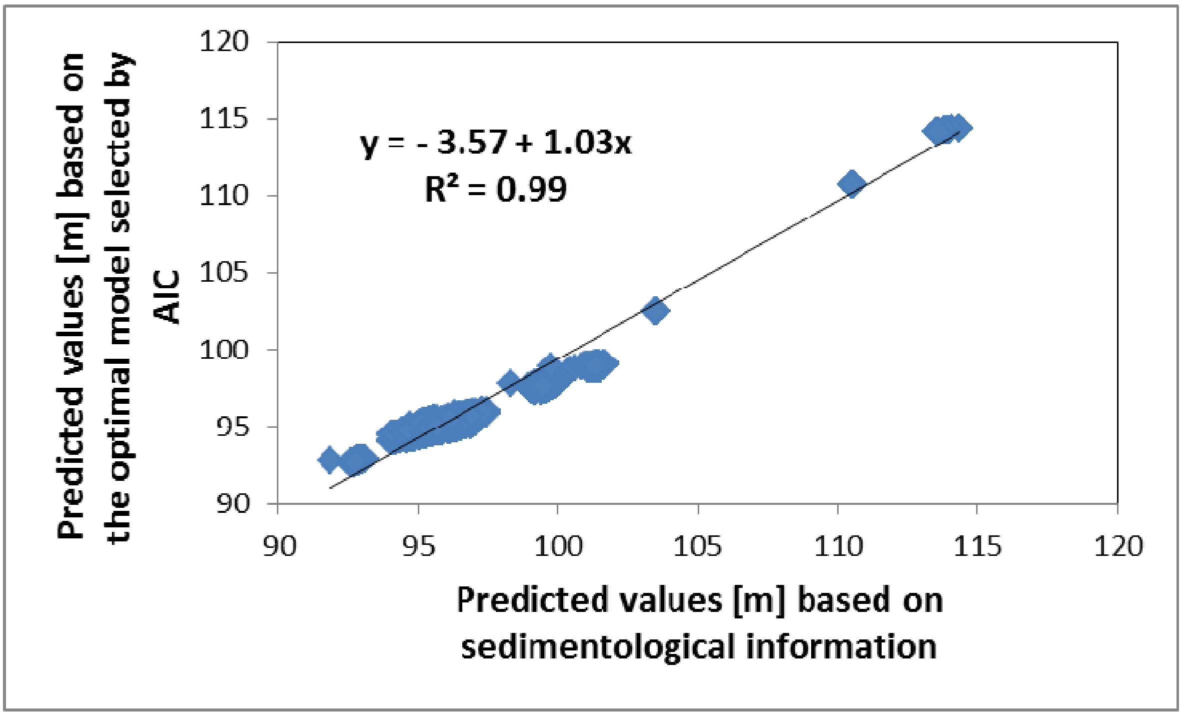

***Fig 8:** Paired model analysis: predicted piezometric pressure heads of Model 1 (based on sedimentological information) versus the results of the optimal model selected by AIC (Model 4), regression line equation, and correlation coefficient (R^2).*

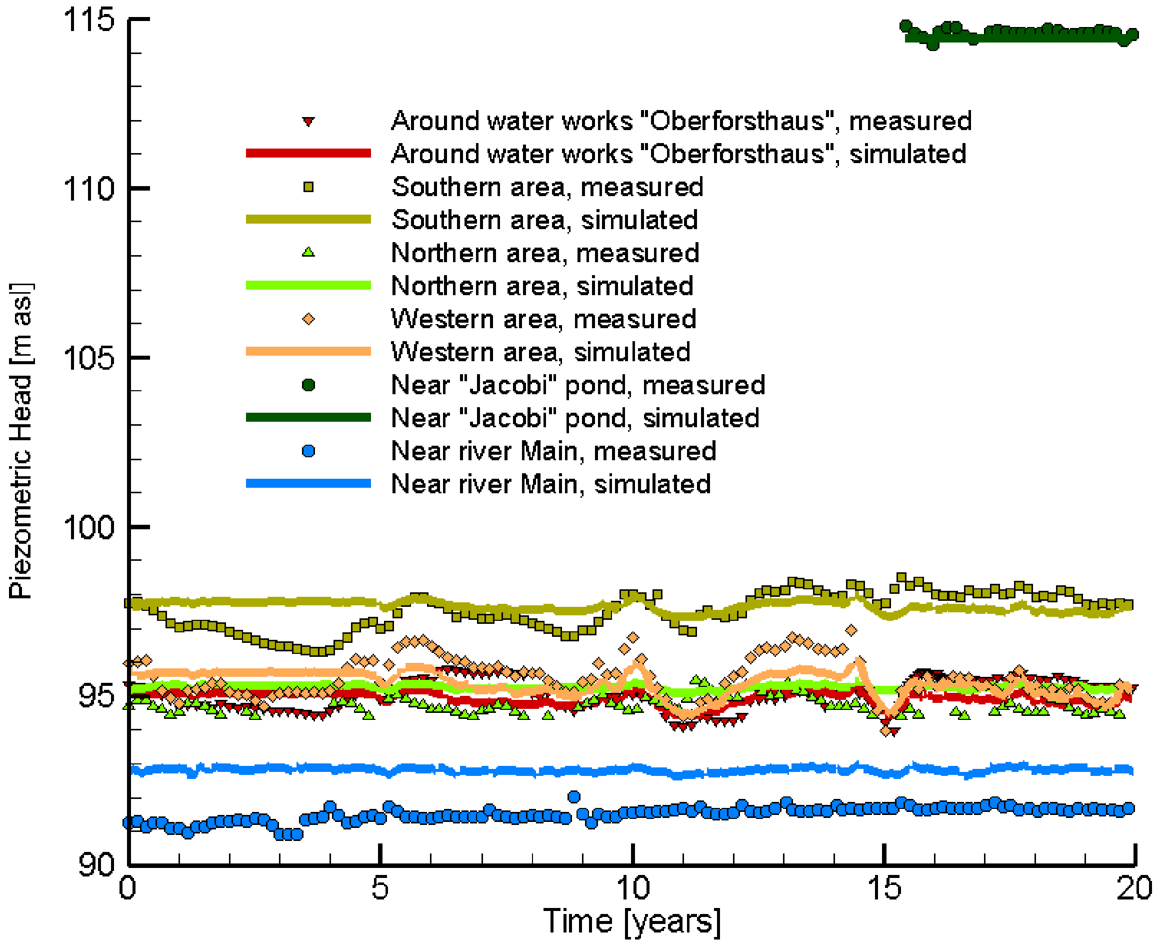

Fig 9: Simulated piezometric heads of Model 4 (optimal model) versus measured piezometric heads between 1990 and 2009. Observation wells were summarized in six groups. One observation well of each group is illustrated within the figure.