Assessing the effects of spatial discretization on large-scale flow model performance and prediction uncertainty

Wildemeersch, S[a,]*, Goderniaux P.[b], Orban Ph.[a], Brouyère S.[a], Dassargues A.[a]

[a] *University of Liege, ArGEnCo, GEO³, Hydrogeology and Environmental Geology, Aquapôle, B52/3 Sart-Tilman, 4000 Liege, Belgium*

[b] *University of Mons, Faculty of Engineering, Fundamental and Applied Geology Department, Rue de Houdain 9, 7000 Mons, Belgium*

* Corresponding author at: University of Liege, ArGEnCo, GEO³, Hydrogeology and Environmental Geology, Aquapôle, B52/3 Sart-Tilman, 4000 Liege, Belgium. Tel.: +32 (0)43669553. Fax: +32 (0)43669520.

E-mail address: swildemeersch@ulg.ac.be (S. Wildemeersch)

Abstract

Large-scale physically-based and spatially-distributed models (>100 km^2) constitute useful tools for water management since they take explicitly into account the heterogeneity and the physical processes occurring in the subsurface for predicting the evolution of discharge and hydraulic heads for several predictive scenarios. However, such models are characterized by lengthy execution times. Therefore, modelers often coarsen spatial discretization of large-scale physically-based and spatially-distributed models for reducing the number of unknowns and the execution times. This study investigates the influence of such a coarsening of model grid on model performance and prediction uncertainty. The improvement of model performance obtained with an automatic calibration process is also investigated. The results obtained show that coarsening spatial discretization mainly influences the simulation of discharge due to a poor representation of surface water network and a smoothing of surface slopes that prevents from simulating properly surface water-groundwater interactions and runoff processes. Parameter sensitivities are not significantly influenced by grid coarsening and calibration can compensate, to some extent, for model errors induced by grid coarsening. The results also show that coarsening spatial discretization mainly influences the uncertainty on discharge predictions. However, model prediction uncertainties on discharge only increase significantly for very coarse spatial discretizations.

Keywords: spatial discretization; model performance; sensitivity analysis; automatic calibration; prediction uncertainty.

1 Introduction

Large-scale physically-based and spatially-distributed models (> 100 km^2) are increasingly used in water management for their unique capacity of gathering every piece of information obtained on a hydrological system to simulate its quantitative and qualitative evolution for several predictive scenarios. These models are intended to provide predictions on both the integrated response (discharge) and the distributed response (hydraulic heads) of the catchment.

Physically-based and spatially-distributed models take explicitly into account the heterogeneity and the physical processes occurring in the surface and the subsurface. Therefore, they are expected to provide predictions with higher level of confidence than black-box models (e.g. Ebel and Loague, 2006; Li et al.; 2008; Goderniaux et al., 2009). Additionally, they are also used for improving the understanding of the physics of hydrological processes (e.g. Frei et al., 2009; Meyerhoff and Maxwell, 2011; Irvine et al., 2012). However, physically-based and spatially-distributed models are characterized by lengthy execution times, especially for integrated surface and subsurface transient flow simulations at large-scale. Consequently, choices and simplifications are made for obtaining tractable execution times. The most common simplification consists in coarsening the spatial discretization for reducing the number of unknowns of the problem and the execution time. The effects of such a coarsening of model grid are worthy being studied since they can limit the accuracy of model results and increase model prediction uncertainties.

A series of studies have already been performed on the effects of spatial discretization on physically-based and spatially-distributed model performance. Refsgaard (1997) calibrated and validated a 3D model with a 500 m grid i.e. with a constant element size of 500 m (no refinement) for the Karup catchment in Denmark (440 km^2). Three other models with 1000 m, 2000 m, and 4000 m grids were then generated using the same parameter values than those obtained by calibration for the initial model (no recalibration, no upscaling). The models were compared in terms of both discharge and hydraulic heads. The results from this study indicated that runoff was poorly simulated by the models coarser than 1000 m due to a poor representation of the surface water network which prevents from simulating properly surface water-groundwater interaction. However, the author suggested that a significant recalibration of models with a coarse grid could improve their performance. This is supported by the study of Vàzquez et al. (2002). They calibrated a 3D model with a 600 m grid i.e. with a constant element size of 600 m (no refinement) for the Gete catchment in Belgium (586 km^2). They also generated a second model with a finer grid (300 m) and a third model with a coarser grid (1200 m) using the same parameters than those obtained by calibration for the initial model (no recalibration, no upscaling). These 300 m and 1200 m grid models were then recalibrated individually using a trial-and-error calibration process. As for the study of Refsgaard (1997), the models were compared in terms of both discharge and hydraulic heads. Although, in general, model results remained worse for the 1200 m grid model than for the 300 m and the 600 m grid models, this study proved that a recalibration is required for obtaining effective parameter values and improving model performance when the grid resolution is changed. Sciuto and Diekkrüger (2010) developed a 3D model with a 25 m grid refined in the river zone for the Wüstebach catchment in Germany (0.27 km^2). They also developed a second model with a 100 m grid using the mean averaging method for upscaling parameter values and a third model with the same model grid than the initial model and the same soil configuration than the second model. They compared the results obtained in terms of discharge and spatial pattern of soil moisture. The influence of upscaling was investigated by comparing the first and the second models and the effects of spatial discretization were studied by comparing the second and the third models. They showed that a coarse grid leads to higher discharge and less actual evapotranspiration than a fine grid due to the smoothing of soil surface which induces a loss of topographic information. They also showed that the upscaling technique they selected was efficient for simulating discharge and spatial pattern of soil moisture. They suggested that the nonlinear relationship between soil moisture and evapotranspiration could explain the deterioration of model results when the grid is coarsened without parameter upscaling. However, none of their models were calibrated. Downer and Ogden (2004) performed a spatial convergence study for the Hortonian Godwin Creek Experimental catchment (21.2 km^2) and the non-Hortonian Muddy Brook catchment (3.64 km^2) in the US. They developed a series of 2D vadose zone model of increasing vertical cell size for each of these catchments. The models were calibrated with an automated calibration process using the shuffled complex evolution method. The calibrated models were compared in terms of infiltration, runoff, and evapotranspiration fluxes to evaluate the appropriate vertical discretization required for accurately solving the Richards' equation. The results from this study showed that small vertical cell size (on the order of centimetres) is required in the unsaturated zone to accurately simulate hydrological fluxes. However, providing that effective parameters obtained by calibration are used, the results of this study also shows that it is possible to slightly increase vertical cell size in the unsaturated zone without significantly deteriorating the simulation of hydrological fluxes. These results about the vertical cell size required in the unsaturated zone for accurately solving the Richards' equation are consistent with those obtained by Vogel and Ippisch (2008).

All these studies provide valuable information on the effects of spatial discretization on model performance. However, most of them neglect the calibration or use a simple trial-and-error calibration process which is, by nature, subjective (Poeter and Hill, 1997). An automatic calibration process is essential for properly evaluating the capacity of calibration to improve the performance of models with a coarse grid. The present study includes such an automatic calibration process. Additionally, the present study includes for the first time an evaluation of the influence of spatial discretization on model prediction uncertainties by comparing the linear confidence intervals on predictions calculated for each model.

The objective of the present study is to evaluate the effects of horizontal spatial discretization on discharge and hydraulic heads simulated by a large-scale physically-based and spatially-distributed model. This evaluation is performed using graphs of model fit and performance criteria. The improvement of model performance obtained with an automatic calibration process is also investigated and linear confidence intervals on predictions are calculated for each model. The results of this study can help modelers defining the horizontal spatial discretization for their models by better perceiving its influence on model performance and model prediction uncertainties.

2 Methodology

The effects of horizontal spatial discretization on model performance and model prediction uncertainties are investigated using a synthetic catchment. The hydrological processes in this synthetic catchment are simulated with HydroGeoSphere (Therrien et al., 2012). HydroGeoSphere is a fully-integrated physically-based hydrological model capable of solving very complex problems such as integrated flow in large-scale catchments (for example, see Goderniaux et al., 2009; 2011). Two-dimensional surface water flow is represented using the two-dimensional diffusion-wave approximation to the Saint-Venant equation. Three-dimensional subsurface water flow in both the saturated and the vadose zones is represented using the Richards' equation. The processes of interception and evapotranspiration are modeled following the conceptualization of Kristensen and Jensen (1975). The coupling of the surface to the subsurface is either performed with the common node approach (continuity of hydraulic head between the two domains) or the dual node approach (exchange of water between the two domains via a first-order exchange coefficient). A complete description of HydroGeoSphere is available in Therrien et al. (2012). A short summary is provided in the paper of Li et al. (2008) and in the software spotlight of Brunner and Simmons (2012a).

The choice of working with a synthetic catchment instead of a real catchment is motivated by the wish of focusing only on the effects of horizontal spatial discretization on model performance. When working with a synthetic catchment, the model geometry, the parameter values, and the boundary conditions are exactly known. Furthermore, there is no measurement error on the observations produced. Therefore, it is possible to test specific model features such as the influence of grid resolution on discharge and hydraulic head simulation without unintentionally taking into account other sources of errors related to a lack of knowledge of the hydrological system. The concept of synthetic catchment is quite usual in hydrogeology (for example, see Poeter and McKenna, 1995; Hill et al., 1998; Schäfer et al., 2004; Bauer et al., 2006, Beyer et al., 2006). The synthetic catchment generated for this study is complex in that the flow system is fully-integrated and physically-based with consistent physical state parameters. However, the synthetic catchment is simplified with respect to the heterogeneity of land use and geology in reality. Yet, this study focuses on the effects of spatial discretization on model performance and not on the influence of heterogeneity representation. The way grid size influences model results would have been similar for a synthetic catchment with a higher level of heterogeneity, provided that the heterogeneity is correctly represented. Therefore, despite this simplification, the synthetic catchment is judged complex enough to serve the objective of this study.

The methodology involves three main steps:

STEP 1 – Generation of the reference model/Generation of models with a coarse grid. A 5-year simulation is run with the reference model for producing reference discharge and hydraulic head observations. The reference model is characterized by a fine spatial discretization. The same 5-year simulation is then run with models with a coarse grid using the same parameter values than those used in the reference model (no calibration). These models with a coarse grid differ by their horizontal spatial discretization (constant element size of 250 m, 500 m, 750 m model, or 1000 m). The simulated values of discharge and hydraulic head obtained with these models are saved for further graphical model fit analysis and calculation of performance criteria.

STEP 2 – Calibration of models with a coarse grid. The models with a coarse grid are individually calibrated using an automatic calibration process in order to evaluate how far parameter values can compensate for errors induced by grid coarsening. However, prior to the calibration, a sensitivity analysis is performed for evaluating the influence of horizontal spatial discretization on parameter sensitivities.

The sensitivity of each parameter included in the calibration process is evaluated using the composite scaled sensitivity css_j (Hill, 1992; Anderman et al., 1996; Hill et al., 1998; Hill and Tiedeman, 2007):

$$css_j = \left[\frac{\sum_{i=1}^{nobs}\left(dss_{ij}\right)^2\Big|_b}{nobs}\right]^{1/2} \quad j = 1, npar \tag{1}$$

with the dimensionless scaled sensitivities dss_{ij} calculated as

$$dss_{ij} = \frac{\partial y_i^{sim}}{\partial b_j}\bigg|_b \times \left|b_j\right| \times w_{ii}^{1/2} \quad i = 1, nobs \quad j = 1, npar \tag{2}$$

The composite scaled sensitivity measures the information provided by the entire set of observations for the estimation of the single parameter b_j. Large values correspond to sensitive parameters for which the observations provide a lot of information. According to Hill et al. (1998) and Hill and Tiedeman (2007), parameters with composite scaled sensitivities less than 1 or less than 0.01 of the largest composite scaled sensitivity are poorly sensitive. Consequently, they could produce problems during the calibration or calibrated parameters with large confidence intervals.

The calibration is performed using PEST (Doherty, 2005) enhanced with the temporary parameter immobilization strategy developed by Skahill and Doherty (2006). The iterative local optimization method implemented in PEST allows calculating the set of parameter values that produces the smallest value of an objective function measuring the discrepancies between observed values and their simulated equivalent. The objective function implemented in PEST is the weighted least-squares objective function (L_2 norm):

$$\Phi(\boldsymbol{b}) = \sum_{i=1}^{nobs} w_i \times \left[y_i^{obs} - y_i^{sim}(\boldsymbol{b})\right]^2 = \sum_{i=1}^{nobs} w_i \times r_i^2 \qquad (3)$$

where $nobs$ is the number of observations of any kind, y_i^{obs} is the ith observed value, $y_i^{sim}(\boldsymbol{b})$ is the simulated equivalent to the ith observed value calculated with the parameter values composing the vector $\boldsymbol{b}$, w_i is the weight for the ith contribution to the objective function, r_i is the ith residual. However, in presence of local minima in the objective function, this method based on local parameter sensitivities does not always provide the set of parameter values corresponding to the global minimum. The use of the temporary parameter immobilization strategy greatly reduces this eventuality. This strategy consists in selectively withdrawing the most insensitive parameters from the estimation process when the objective function improvement during a particular iteration is poor. This greatly heightens the capacity of the estimation process to find the global minimum of the objective function. According to Doherty

(2005), calibration using truncated singular value decomposition, gives similar results since this method also has the capacity of withdrawing insensitive parameters from the estimation process. Global optimization methods ensuring to find the global minimum of the objective function are not used in this study because they require a huge number of model runs which induces execution times tens or hundreds of times longer than the execution times required by local optimization methods (Hill and Tiedeman, 2007). This precludes using these methods for integrated surface and subsurface transient flow simulations at large-scale due to their long execution times.

The set of parameters included in both the sensitivity analysis and the automatic calibration is composed of 32 parameters corresponding to the physical state parameters found in the equations representing surface and subsurface flow processes in HydroGeoSphere. The parameters present in the equations representing the interception and evapotranspiration processes are not included. The set of observations is composed of 24 discharge rates and 288 hydraulic head observations (1 per month and per observation point for 2 years) produced with the synthetic catchment. The simulated values of discharge and hydraulic head obtained with these calibrated models are saved for further graphical model fit analysis and calculation of performance criteria.

STEP 3 – Graphics of model fit, performance criteria, and linear confidence intervals on predictions. Graphical model fit analysis and calculation of performance criteria are performed for each model to evaluate qualitatively and quantitatively the effects of spatial discretization on model performance and to evaluate the improvement of model performance obtained with calibration. Additionally, the influence of horizontal spatial discretization on model prediction uncertainties is evaluated using linear confidence intervals on predictions.

Graphics of model fit. Graphical model fit analysis is somewhat subjective. However, it is good practice to perform such a visual inspection prior to use numerical criteria for an objective evaluation of model performance (Legates and McCabe, 1999; Hill and Tiedeman, 2007; Moriasi et al, 2007). Graphs comparing observed and simulated values are the most widely used for evaluating model fit at a glance. However, Hill and Tiedeman (2007) prefer using graphs such as weighted or unweighted simulated values versus weighted residuals to facilitate the detection of model bias. If a model is unbiased, such graphs exhibit weighted residuals evenly scattered about 0.0 for the entire range of values on the horizontal axis. Weighted residuals wr_i are calculated as (Hill and Tiedeman, 2007):

$$wr_i = w_i^{1/2} \times \left[y_i^{obs} - y_i^{sim}\right] = w_i^{1/2} \times r_i \qquad (4)$$

The purpose of weighting is essentially to emphasize the most accurate observations. This is achieved by specifying weights that are proportional or, preferably, equal to the inverse of the observation error variances (Hill and Tiedeman, 2007):

$$w_i = \frac{1}{\sigma_i^2} \qquad (5)$$

where σ_i^2 is the true error variance of the i^{th} observation. Given these equations, in a graph of weighted residuals versus unweighted simulated values, a cluster of negative weighted residuals indicate that simulated values are systematically overestimated, and vice versa. Furthermore, with weights calculated using a constant coefficient of variation, residuals are emphasized proportionally to their observed value. Therefore, similar weighted residuals indicate similar relative errors. This way of emphasizing residuals proportionally to their observed value is particularly useful for variables ranging over several orders of magnitudes such as discharge.

Performance criteria. Performance criteria help quantifying model quality. They evaluate the level of agreement between model and reality (Refsgaard and Henriksen, 2004). Typically, they depend on the discrepancies between observed values and their simulated equivalent for a particular type of observations (e.g. discharge or hydraulic heads). The performance criteria selected for this study are:

- The Nash-Sutcliffe efficiency criterion NSE_q (Nash and Sutcliffe, 1970):

$$NSE_q = 1 - \frac{\sum_{t=1}^{nt}\left(q_t^{sim} - q_t^{obs}\right)^2}{\sum_{t=1}^{nt}\left(q_t^{sim} - \mu^{obs}\right)^2} \in]-\infty; 1] \qquad (6)$$

where nt is the total number of timesteps, q_t^{sim} is the simulated discharge at timestep t, q_t^{obs} is the observed discharge at timestep t, and μ^{obs} is the mean of the observed values. If the simulated values perfectly match the observed values, $NSE_q = 1$. The lower the value of NSE_q, the poorer the model, negative values indicating that the mean observed value μ^{obs} gives a better description of the data than the simulated values q_t^{sim}. Weglarczyk (1998) and Gupta et al. (2009) suggest decomposing the Nash-Sutcliffe efficiency criterion for facilitating its interpretation. The decomposition of Gupta et. al (2009) is:

$$NSE_q = 2 \times \frac{\sigma^{sim}}{\sigma^{obs}} \times r_{lin} - \left(\frac{\sigma^{sim}}{\sigma^{obs}}\right)^2 - \left(\frac{\mu^{sim} - \mu^{obs}}{\sigma^{obs}}\right)^2 \qquad (7)$$

where r_{lin} is the linear correlation coefficient between q^{sim} and q^{obs}, σ^{sim} is the standard deviation of q^{sim}, μ^{sim} is the mean of q^{sim}, σ^{obs} is the standard deviation of q^{obs}, and μ^{obs} is the mean of q^{obs}. The first component uses the linear correlation coefficient for measuring the capacity of the model to reproduce the timing and the shape of the signal, the second component measures the capacity of the model to reproduce the standard deviations of the observations, and the third component measures the capacity of the model to reproduce the mean of the observations.

- The mass balance error MBE_q also known as bias, percent bias or relative bias (Gupta et al., 1999):

$$MBE_q = \frac{\sum_{t=1}^{nt}\left(q_t^{sim} - q_t^{obs}\right)}{\sum_{t=1}^{nt} q_t^{obs}} \times 100 = \frac{\mu^{sim} - \mu^{obs}}{\mu^{obs}} \times 100 \in]-100; +\infty[\qquad (8)$$

This performance criterion measures the tendency of the simulated values to be larger or smaller than their observed counterparts. If the fit is perfect, $MBE_q = 0$. If $MBE_q > 0$, simulated values are, on average, greater than observed values, and vice versa. This performance criterion can also be used for hydraulic heads by substituting the observed and simulated discharges by the observed and simulated hydraulic heads in equation (7).

- The peak error PE_q (Aricò et al., 2009):

$$PE_q = \left(\frac{q_{peak}^{sim}}{q_{peak}^{obs}} - 1\right) \times 100 \;\in]-100; +\infty[\qquad (9)$$

where q_{peak}^{sim} is the simulated peak value, and q_{peak}^{obs} is the observed peak value. This performance criterion measures the capacity of the model to reproduce the peak in the hydrograph. If the observed peak is equal to the simulated peak, $PE_q = 0$. If $PE_q > 0$, the simulated peak is greater than the observed peak, and vice versa.

- The root mean squared error criterion RMS_h:

$$RMS_h = \sqrt{\frac{1}{nt} \times \sum_{t=1}^{nt}\left(h_t^{sim} - h_t^{obs}\right)^2} \;\in [0; +\infty[\qquad (10)$$

where h_t^{sim} is the i^{th} simulated hydraulic head value, and h_t^{obs} is the i^{th} observed hydraulic head value. This performance criterion measures the discrepancies between observed hydraulic heads and their simulated equivalent for a particular observation point. If the simulated values perfectly match the observed values, $RMS_h = 0$. The greater the values, the poorer the model.

- The hydraulic head variations errors $HHVE_h$:

$$HHVE_h = \left(\frac{h_{max}^{sim} - h_{min}^{sim}}{h_{max}^{obs} - h_{min}^{obs}} - 1\right) \times 100 \;\in]-100; +\infty[\qquad (11)$$

where h_{max}^{sim} is the maximum simulated hydraulic head value, h_{min}^{sim} is the minimum simulated hydraulic head value, h_{max}^{obs} is the maximum observed hydraulic head value, and h_{min}^{obs} is the minimum observed hydraulic head value. This performance criterion is the counterpart of the peak error since it measures the capacity of the model to reproduce the magnitude of hydraulic head variations instead of measuring the capacity of the model to reproduce the peak in the hydrograph.

***Linear confidence intervals*.** Linear and nonlinear confidence intervals help quantifying prediction uncertainties. Linear confidence intervals are calculated assuming that the model is linear in the vicinity of parameter values. They are not as accurate as nonlinear confidence intervals for nonlinear models. However, unlike nonlinear confidence intervals, linear confidence intervals only require trivial amount of execution time. Therefore, they are often the only confidence intervals calculable for physically-based and spatially-distributed models with lengthy execution times.

Linear confidence intervals on predictions have the form:

$$z_l' \pm t_S(n, 1.0 - \tfrac{\alpha}{2}) \times s_{z_l'} \qquad (12)$$

where z_l' is the lth simulated prediction, $t_S\left(n, 1.0 - \frac{\alpha}{2}\right)$ is the Student-t distribution with $n = (nobs - npar)$ and $\alpha = 0.05$ for 95% confidence intervals, and $s_{z_l'}$ is the standard deviation of the prediction calculated as:

$$s_{z_l'} = \left[\sum_{i=1}^{npar} \sum_{j=1}^{npar} \frac{\partial z_l'}{\partial b_j} \times \boldsymbol{V}(\boldsymbol{b}) \times \frac{\partial z_l'}{\partial b_i}\right]^{1/2} \qquad (13)$$

where $npar$ is the number of parameters, $\frac{\partial z_l'}{\partial b_j}$ is the sensitivity of the lth prediction z_l' with respect to the jth parameter b_j and $\boldsymbol{V}(\boldsymbol{b})$ is the parameter variance-covariance matrix.

3 Conceptual model

The synthetic catchment is inspired by a real catchment located in the Condroz region of Belgium. This region is characterized by a succession of limestone synclines and sandstone anticlines. The surface materials of the synthetic catchment are assigned using a criterion combining elevation and slope constraints. All surface materials are assigned a series of parameters required for simulating interception, evapotranspiration, and surface flow processes. Appropriate values for these parameters are extracted from the literature. They are listed in Appendix A. The subsurface materials of the synthetic catchment are defined to represent the typical geology of the Condroz region: sandstones, limestones, and shales constitute, from the crests to the center of the valley, the subsurface materials of the synthetic catchment. Additionally, these formations are covered by alluvial deposits and loam. All subsurface materials are assigned a series of parameters required for simulating subsurface flow processes. Appropriate values for these parameters, including van Genuchten parameters governing saturation-pressure relations in the vadose zone, are extracted from the literature. They are listed in Appendix B. The synthetic catchment is illustrated in Figure 1.

The horizontal element size of the reference model progressively increases from 25 m near the surface water network to 250 m far from the surface water network. The layer thickness progressively increases from 1 m for the top layers corresponding to the vadose zone to 30 m for the bottom layers corresponding to the saturated zone (5 layers of 1 m, 1 layer of 5 m, 1 layer of 10 m, and 1 layer of 30 m). The reference model is composed of 153,027 nodes and 269,872 elements. The grid of the reference model is illustrated in Figure 2. Critical-depth boundary conditions are assigned to boundary nodes of the surface domain. This type of boundary condition forces the water elevation at the boundary to be equal to the water elevation for which the energy of the flowing water relatively to the stream bottom is minimum (Therrien et al., 2012). No-flow boundary conditions are assigned to boundary nodes of the subsurface domain. Water depths and hydraulic heads extracted from preliminary simulations performed with the reference model are used as initial conditions for the surface domain and the subsurface domain, respectively.

The set of observation points is constituted of 1 gauging station for discharge (G1) and 12 piezometers evenly distributed in the synthetic catchment for hydraulic heads (Pz1 to Pz12). Two galleries (GAL1 and GAL2) and four wells (W1 to W4) are used to simulate groundwater withdrawals. The set of observation points and the galleries and wells are illustrated in Figure 1. As the models with a coarse grid are run with monthly stress factors, discharge and hydraulic heads simulated at the observation points each day of the 5-year reference simulation are monthly averaged for ensuring time consistency (Hill and Tiedeman, 2007, p. 215). These monthly averaged discharge and hydraulic heads constitute the set of reference observations used to calculate performance criteria for the simplified models. The reference simulation is subdivided into warm-up, calibration, and validation periods. The warm-up is necessary for obtaining simulated values independent of the initial conditions. Discharge and hydraulic heads produced during the warm-up period are not included in the set of reference observations. Performance criteria are only calculated for discharge and hydraulic heads produced during calibration and validation periods. Linear confidence intervals on predictions are calculated for the validation period.

4 Results and Discussion

The models developed for evaluating the effects of spatial discretization on model performance and model prediction uncertainties are referred as the *250 m*, *500 m*, *750 m*, and *1000 m* models. They are characterized by a constant element size of 250 m, 500 m, 750 m, and 1000 m, respectively. As opposed to the reference model, they are not refined near the surface water network. The purpose here consists in evaluating the effects of ignoring such a refinement on the simulation of discharge and hydraulic heads. Additionally, it also allows evaluating whether calibration can compensate for the errors induced by ignoring such a refinement. As the reference model, they each have 8 layers (5 layers of 1 m, 1 layer of 5 m, 1 layer of 10 m, and 1 layer of 30 m). The number of nodes, the number of elements, and the execution times of the *250 m*, *500 m*, *750 m*, and *1000 m* models are presented in Table 1. The comparison between the execution time of each model clearly shows the usefulness of coarsening grid size for reducing the execution times.

4.1 Comparison of model performance before calibration

Graphs of model fit and performance criteria are used together for comparing the performance of the *250 m*, *500 m*, *750 m*, and *1000 m* models run with the same parameter values than those used in the reference model i.e. without any calibration.

Graphs comparing reference values of discharge and hydraulic heads produced with the reference model and their simulated equivalent obtained with the models with a coarse grid indicate that discharge is most often underestimated during low flow periods and overestimated during high flow periods (Figure 3-A). The underestimation is almost identical for each model. The overestimation is higher for models with a coarse horizontal spatial discretization. This is clearly visible on peak discharge.

Graphs of unweighted simulated values versus weighted residuals support these findings. These graphs particularly highlight the underestimation of discharge by each model during low flow periods and the overestimation of discharge during high flow rates by the coarsest ones (Figure 3-B).

The influence of horizontal spatial discretization on hydraulic head simulation is less visible (Figure 3-A). However, weighted residuals are in general greater for the coarsest models (Figure 3-B). This shows that the simulation of hydraulic heads is poorer with the coarsest models.

Graphical model fit analysis is confirmed by performance criteria. As the grid is coarsened, NSE_q values tend to decrease and RMS_h values tend to increase (Figure 4-A). This indicates that simulation of both discharge and hydraulic heads is deteriorated. For discharge, Gupta's decomposition of NSE_q shows that the standard deviation of discharge is overestimated by the coarsest models (Table 2). This is visible to the greater values of Gupta's second terms. This is also supported by the increasing values of $PE_q^{yr\,1}$ and $PE_q^{yr\,2}$ showing that peak discharge, and so the standard deviation of the hydrograph, are overestimated by the coarsest models (Table 2). Gupta's decomposition also shows that the *250 m* model lacks to properly simulate the average magnitude of discharge. This is why NSE_q value for this model is lower than NSE_q value for the *500 m* model. This is confirmed by the values of MBE_q which shows that the *250 m* model underestimates the average magnitude of discharge by almost 15%. This is related to the fact that the underestimation of discharge during low flow periods is not compensated by the overestimation of discharge during high flow rates as it is the case for the other models. For hydraulic heads, the absolute values of MBE_h are in general low for each model (Table 3). This indicates that models are not significantly biased in terms of hydraulic heads. However, the range of MBE_h values is in general wider for the coarsest models. Although the absolute values of $HHVE_h^{yr\,1}$ and $HHVE_h^{yr\,2}$ are in general greater for the coarsest models, the ranges of $HHVE_h^{yr\,1}$ and $HHVE_h^{yr\,2}$ are similar for each model (Table 3).

The comparison of model performance performed in this section indicates that coarsening the grid mainly deteriorates the simulation of discharge. Common to each model tested, the underestimation of discharge during low flow periods is due to a poor representation of the surface water network which precludes from properly simulating groundwater-surface water interactions that constitute the key component of the hydrograph during dry seasons. As previously mentioned, this problem of poor representation of the surface water network is also mentioned by Refsgaard (1997) and Vázquez et al. (2002). The overestimation of discharge by the coarsest models during high flow periods is related to the use of large elements which induces a smoothing of surface slopes and facilitates runoff, especially during wet seasons. The object of the next section is to evaluate how calibration can compensate for the errors induced by coarsening the grid.

4.2 Comparison of model performance after calibration

A sensitivity analysis is performed for each parameter prior to the calibration. The composite scaled sensitivities calculated on the calibration period (24 discharge observations and 288 hydraulic head observations) for each parameter included in the calibration (32 parameters) are illustrated in Figure 5. Whatever the spatial discretization, the ranking of the most sensitive parameters and the magnitude of the composite scaled sensitivities are almost identical. This suggests that parameter sensitivities are not highly dependent on the grid size. The most sensitive parameter is always the van Genuchten parameters β_{vG} of Mat I – loam and Mat II – alluvial deposits (top layers of the models). This parameter, related to the pore-size distribution in the porous medium, defines the shape of the water retention curve. The other most sensitive parameters are the hydraulic conductivity K of Mat IV – limestones 2, probably because most of the observation points are located in this material, and the van Genuchten parameter β_{vG} of Mat IV – limestones 1, Mat V – limestones 2 and Mat VI – sandstones. The van Genuchten parameter α_{vG} of Mat IV – limestones 1, Mat V – limestones 2 and Mat VI – sandstones as well as the hydraulic conductivity K of Mat I – loam have also a relatively high sensitivity. The fact that van Genuchten parameters, especially the parameter β_{vG} of the materials constituting the top layers of the models, are systematically among the most sensitive parameters suggests that fully-integrated and physically-based models are highly sensitive to parameters governing the infiltration process in the vadose zone and the groundwater recharge.

The improvement of model performance with calibration with PEST is evaluated using the same graphs of model fit and the same performance criteria than in the previous section. Graphs of model fit show that calibration significantly improves the simulation of discharge and, to a lesser extent, hydraulic heads for each model (Figure 6-A). Additionally, after calibration,weighted residuals are almost randomly distributed which suggests that calibrated models are less biased (Figure 6-B). Performance criteria support these findings since NSE_q and RMS_h values are significantly greater and lower, respectively, after calibration (Figure 4-B). The values of Gupta's terms together with the values of MBE_q, $PE_q^{yr\,1}$, and $PE_q^{yr\,2}$ calculated for the calibrated models indicate that both the mean and the standard deviation of flow rates are better simulated (Table 4). The improvement of hydraulic head simulation is not so clear. When observed and simulated hydraulic heads are shifted, the calibration process strives for reducing this systematic error. Therefore, the improvement of average hydraulic head magnitudes is sometimes obtained to the detriment of the improvement of hydraulic head variations. This is why the absolute values and the range of MBE_h are most often lower than those obtained with the models before calibration, while the absolute values and the ranges of $HHVE_h^{yr\,1}$ and $HHVE_h^{yr\,2}$ are identical or even greater than those obtained with the models before calibration (Table 5). This shows that calibration has limitations. Furthermore, although most of them are still within reasonable ranges, some calibrated parameter values are far from their values in the reference model (Table 6). Such an observation is only possible for synthetic catchments for which reference parameter values are exactly known. The only verification possible for real catchments consists in making sure that calibrated parameter values are plausible with regards to field or laboratory data. However, as shown by Brunner et al. (2012b), accurately evaluating certain combinations of parameters can be sufficient to produce predictions with a good level of confidence, which means that it is not always necessary to accurately evaluate each parameter individually. Therefore, in spite of its limitations, calibration is essential for improving model performance, either inside or outside the calibration period. As illustrated in Figure 7, calibration indeed leads to greater values of NSE_q and lower values of RMS_h also during the validation period. The object of the next section is to evaluate whether grid coarsening leads to greater model prediction uncertainties.

4.3 Comparison of model prediction uncertainties

Linear confidence intervals on predictions are calculated for discharge and hydraulic heads simulated in the validation period with the calibrated models. They are illustrated in Figures 8 to 11. The linear confidence intervals calculated for discharge are almost identical for the *250 m*, *500 m*, and *750 m* models. They are even sometimes narrower for the *500 m* or the *750 m* models than for the *250 m* model. However, especially for high flow periods, they are far wider for the *1000 m* model. The linear confidence intervals calculated for hydraulic heads are quite similar for each model and once more the narrowest intervals are not always obtained for the *250 m* model.

The analysis of model prediction uncertainties indicates that coarsening model grid mainly influences the uncertainties on discharge predictions. This is not surprising since the comparison of model performance shows that discharge simulation is more sensitive to grid size than hydraulic head simulation. However, the uncertainties on discharge predictions significantly increase only for a very coarse grid and even if graphs of model fit and performance criteria suggest that the model is good. Therefore, to some extent, it is possible to simplify a model by coarsening its grid without increasing model prediction uncertainties. This is consistent with the study of Brunner et al. (2012b) focusing on parameter identifiability and predictive uncertainty. This study highlights the sliding nature of complexity versus simplicity and shows that predictive power may lose little if the model is simplified appropriately.

4.4 Guidelines for selection of a proper horizontal spatial discretization for large-scale flow modelsA synthetic catchment can always be considered as far from reality. Therefore, caution should be exercised when using results of this study for selecting a proper horizontal spatial discretization for a given site-specific study. However, a series of general guidelines can be drawn from this study. As an example, in the framework of use of paired simple and complex models to reduce predictive bias and quantify uncertainty (Doherty and Christensen, 2011), these guidelines could be used for helping modelers selecting a proper horizontal spatial discretization for the simple model.

Large-scale physically-based and spatially-distributed model development consists in finding a compromise between model accuracy and model portability i.e. maximizing model performance and minimizing prediction uncertainty while limiting the execution times. Given the results of this study, for catchments of a few hundreds square kilometer, an element size of the order of 500 m is the best compromise for obtaining good model performance with tractable execution times without significantly increasing prediction uncertainty. With a coarser horizontal spatial discretization, the relative reduction of execution times is limited with respect to the probability of increasing prediction uncertainty. With a finer horizontal spatial discretization, the execution times strongly increase without any significant reduction of prediction uncertainty.

5 Summary and Conclusions

The present study focuses on the effects of horizontal spatial discretization on large-scale flow model performance and model prediction uncertainties using a fully-integrated hydrological model of a synthetic catchment. This kind of large-scale fully-integrated hydrological model is increasingly used in water management for predicting the evolution of both the integrated response (discharge) and the distributed response (hydraulic heads) of catchments. However, these models are characterized by lengthy execution times and model grids are often coarsened for reducing these execution times. Therefore, it is crucial to evaluate the influence of such a grid coarsening on model performance and model prediction uncertainties. This study shows that:

- Grid coarsening mainly influences the simulation of discharge with an underestimation of discharge during low flow periods and a progressive overestimation of peak discharge as horizontal spatial discretization is coarsened. This is related to a poor representation of the surface water network and the smoothing of surface slopes that prevent from properly simulating surface water-groundwater interactions and runoff process.
- Parameter sensitivities are not significantly influenced by grid coarsening and model errors induced by grid coarsening can be compensated by calibration (preferably using an automatic calibration process). Furthermore, calibration improves model performance either inside or outside the calibration period. However, calibration has limitations and model errors are potentially compensated at the cost of less plausible parameter values.
- Grid coarsening mainly influences the uncertainty on discharge predictions. However, model prediction uncertainties on discharge only increase significantly for very coarse horizontal spatial discretizations.

As uncertainty analyses have become essential in natural system modeling, this is encouraging since grid coarsening greatly reduces execution times and such analyses can only be performed for model with relatively short execution times.

References

Anderman, E., Hill, M.C., Poeter, E.P., 1996. Two-dimensional advective transport in ground-water flow parameter estimation. Ground Water 34(6), 1001-1009.

Aricò, C., Nasello, C., Tucciarelli, T., 2009. Using steady-state water level data to estimate channel roughness and discharge hydrograph. Advances in Water Resources 32(8), 1223-1240.

Bauer, S., Beyer, C., Kolditz, O., 2006. Assessing measurement uncertainty of first-order degradation rates in heterogeneous aquifers. Water Resources Research 42(W01420).

Beyer, C., Bauer, S., Kolditz, O., 2006. Uncertainty assessment of contaminant plume length estimates in heterogeneous aquifers. Journal of Contaminant Hydrology 87(1-2), 73-95.

Brunner, P., Doherty, J., Simmons, C.T., 2012a. Uncertainty assessment and implications for data acquisition in support of integrated hydrologic models. Water Resources Research 48(W07513).

Brunner, P., Simmons, C.T., 2012b. HydroGeoSphere : A fully-integrated, physically-based hydrological model. Ground Water 50(2), 170-176.

Doherty, J., 2005. PEST – Model-independent parameter estimation – User Manual – 5th edition. Watermark Numerical Computing.

Dohery, J., Christensen, S., 2011. Use of paired simple and complex models to reduce predictive bias and quantify uncertainty. Water Resources Research 47(W12534).

Downer, C.W., Ogden, F.L., 2004. Appropriate vertical discretization of Richards' equation for two-dimensional watershed-scale modelling. Hydrological Processes 18, 1-22.

Ebel, B., Loague, K., 2006. Physics-based hydrologic-response simulation: seeing through the fog of equifinality. Hydrological Processes 20(13), 2887-2900.

Frei, S., Fleckenstein, J.H., Kollet, S.J., Maxwell, R.M., 2009. Patterns and dynamics of river-aquifer exchange with variably-saturated flow using a fully-coupled model. Journal of Hydrology 375, 383-393.

Goderniaux, P., Brouyère, S., Fowler, H., Blenkinsop, S., Therrien, R., Orban Ph., Dassargues, A., 2009. Large scale surface-subsurface hydrological model to assess climate change impacts on groundwater reserves. Journal of Hydrology 373(1-2), 122-138.

Goderniaux, P., Brouyère, S., Blenkinsop, S., Burton, A., Fowler, H., Orban Ph., Dassargues, A., 2011. Modeling climate change impacts on groundwater resources using transient stochastic climatic scenarios. Water Resources Research 47 (W12516).

Gupta, H., Sorooshian, S., Yapo, P., 1999. Status of automatic calibration for hydrologic models: comparison with multilevel expert calibration. Journal of Hydrologic Engineering 4(2), 135-143.

Gupta, H., Kling, H., Yilmaz, K., Martinez, G., 2009. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. Journal of Hydrology, 377(1-2), 80-91.

Hill, M.C., 1992. A computer program (MODFLOWP) for estimating parameters of a transient, three-dimensional, ground-water flow model using nonlinear regression. Open-File Report 91-484, U.S. Geological Survey.

Hill, M.C., Cooley, R., Pollock, D., 1998. A controlled experiment in groundwater flow model calibration using nonlinear regression. Ground Water 36(3), 520-535.

Hill, M.C., Tiedeman, C.R., 2007. Effective groundwater model calibration with analysis of sensitivities, predictions and uncertainty. Hoboken, NJ, USA: John Wiley & Sons.

Irvine, D.J., Brunner, P., Hendricks Franssen, H.-J., Simmons C.T., 2012. Heterogeneous or homogeneous? Implications of simplifying heterogeneous streambeds in models of losing streams. Journal of Hydrology 424-425, 16-23.

Kristensen, K.J., Jensen, S.E., 1975. A model for estimating actual evapotranspiration from potential evapotranspiration. Nordic Hydrology 6, 170-188.

Legates, D., McCabe, G., 1999. Evaluating the “goodness-of-fit” measures in hydrologic and hydroclimate model validation. Water Resources Research 35(1), 233-241.

Li, Q., Unger, A., Sudicky, E., Kassenaar, D., Wexler, E., Shikaze, S., 2008. Simulating the multi-seasonal response of a large-scale watershed with a 3D physically-based hydrologic model. Journal of Hydrology 357(3-4), 317-336.

Meyerhoff, S.B., Maxwell, R.M., 2011. Quantifying the effects of subsurface heterogeneity on hillslope runoff using a stochastic approach. Hydrogeology Journal 19, 1515-1530.

Moriasi, D., Arnold, J., Van Liew, M., Binger, R., Harmel, R., Veith, T., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE 50(3), 885-900.

Nash, J., Sutcliffe, J., 1970. River flow forecasting through conceptual models I – A discussion of principles. Journal of Hydrology 10(3), 282-290.

Poeter, E.P., McKenna, S., 1995. Reducing uncertainty associated with ground-water flow and transport predictions. Ground Water 33(6), 899-904.

Poeter, E.P., Hill, M.C., 1997. Inverse models: a necessary next step in ground-water modeling. Ground Water 35(2), 250-260.

Refsgaard, J.C, 1997. Parameterisation, calibration and validation of distributed hydrological models. Journal of Hydrology 198, 69-97.

Refsgaard, J.C., Henriksen, H., 2004. Modelling guidelines – terminology and guiding principles. Advances in Water Resources 27(1), 71-82.

Schäfer, D., Schlenz, B., Dahmke, A., 2004. Evaluation of exploration and monitoring methods for verification of natural attenuation using the virtual aquifer approach. Biodegradation 15(6), 453-465.

Sciuto, G., Diekkrüger, B., 2010. Influence of soil heterogeneity and spatial discretization on catchment water balance modeling. Vadose Zone Journal 9, 955-969.

Skahill, B., Doherty, J., 2006. Efficient accommodation of local minima in watershed model calibration. Journal of Hydrology 329(1-2), 122-139.

Therrien, R., McLaren, R., Sudicky, E., Park, Y.-J.., 2012. HydroGeoSphere – A three-dimensional numerical model describing fully-integrated subsurface and surface flow and solute transport. Manual, Groudnwater Simulations Group.

Vázquez R.F., Feyen, L., Feyen, J., Refsgaard, J.C., 2002. Effect of grid size on effective parameters and model performance of the MIKE-SHE code. Hydrological Processes 16, 355-372.

Weglarczyk, S., 1998. The interdependence and applicability of some statistical quality measures for hydrological models. Journal of Hydrology 206(1-2), 98-103.

Appendices

Appendice A: Interception, evapotranspiration and surface flow parameters used in the synthetic case

Interception and evapotranspiration	**Mat 1**	**Mat 2**	**Mat 3**	**Mat 4**
LAI [-]	0.40	3.53	5.12	-
L_r [m]	0	2.30	2.90	-
L_e [m]	2.00	2.00	2.00	-
θ_{e2} & θ_{t2} [-]	0.60	0.60	0.60	-
θ_{e1} & θ_{t1} [-]	0.96	0.96	0.96	-
C_1 [-]	0.31	0.31	0.31	-
C_2 [-]	0.15	0.15	0.15	-
C_3 [-]	10.00	10.00	10.00	-
C_{int} [m]	5.00×10^{-5}	5.00×10^{-5}	5.00×10^{-5}	-
S^0_{int} [m]	0	0	0	-
Surface flow	**Mat 1**	**Mat 2**	**Mat 3**	**Mat 4**
n_{xx} & n_{yy} [$m^{-1/3}s$]	0.012	0.200	0.600	0.025
H_{sto} [m]	0.002	0.002	0.002	0.002
L_c [m]	1.00×10^{-1}	1.00×10^{-1}	1.00×10^{-1}	1.00×10^{-1}

LAI = Leaf Area Index; L_r = root depth; L_c = evaporation depth; θ_{e2} & θ_{t2} and θ_{e1} & θ_{t1} = evaporation and transpiration limiting saturations; C_1, C_2, and C_3 = transpiration fitting parameters; C_{int} = canopy storage parameter; S^0_{int} = initial interception storage; n_{xx} & n_{yy} = Manning roughness coefficients; H_{sto} = rill storage height; L_c = coupling length.

Parameter values are extracted from the literature:

- for parameters related to interception and evapotranspiration processes, see Andersen et al., 2002; Asner et al., 2003; Canadell et al., 1996; Dickinson et al., 1991; Goderniaux, 2010; Graham and Kilde, 2002; Islam, 2004; Kristensen and Jensen, 1975;Li et al., 2008; Panday and Huyakorn, 2004; Schroeder et al., 2004; Therrien et al., 2005; Vázquez et al., 2002; Vázquez and Feyen, 2003.
- for parameters related to surface flow processes, see Brutsaert, 2005; Fetter, 2001; Hornberger et al., 1998; Jones, 2005; Li et al., 2008; McCuen, 1989.

Appendice B: Subsurface flow parameters used in the synthetic case

Subsurface flow	**Mat I**	**Mat II**	**Mat III**	**Mat IV**	**Mat V**	**Mat VI**
K [ms^{-1}]	5.00×10^{-7}	1.00×10^{-6}	1.00×10^{-5}	1.00×10^{-4}	2.50×10^{-4}	5.00×10^{-5}
S_S [m^{-1}]	1.00×10^{-4}	1.00×10^{-4}	1.00×10^{-4}	1.00×10^{-4}	1.00×10^{-4}	1.00×10^{-4}
θ_s [-]	4.10×10^{-1}	4.10×10^{-1}	2.50×10^{-2}	1.00×10^{-1}	1.00×10^{-1}	7.50×10^{-2}
S_{wr} [-]	9.76×10^{-2}	9.76×10^{-2}	0	0	0	0
α_{vG} [m^{-1}]	2.67	2.67	6.08×10^{-12}	3.65×10^{-2}	3.65×10^{-2}	3.65×10^{-2}
β_{vG} [-]	1.45	1.45	0.62	1.83	1.83	1.83
γ_{vG} [-]	$1-1/\beta_{vG}$	$1-1/\beta_{vG}$	38,671.00	$1-1/\beta_{vG}$	$1-1/\beta_{vG}$	$1-1/\beta_{vG}$

K = saturated hydraulic conductivity; S_S = specific storage; θ_s = saturated water content; S_{wr} = residual water saturation; α_{vG} , β_{vG}, and γ_{vG} = van Genuchten parameters.

Parameter values are extracted from the literature:

- for parameters related to subsurface flow processes, see Brouyère et al., 2009; Freeze and Cherry, 1979; Jones, 2005; Radcliffe, 2000; Ramos da Silva et al., 2008; Roulier et al., 2006.

Figure captions

Figure 1 The reference model is assigned surface materials depending on elevation and slope constraints and subsurface materials following the typical syncline structure of catchments located in the Condroz region of Belgium. A gauging station (G1) and twelve piezometers (Pz1 to Pz12) are used to obtain reference observations in terms of discharge and hydraulic heads, respectively. Two galleries (GAL1 and GAL2) and four wells (W1 to W4) are used to simulate groundwater withdrawals.

Figure 2 The grid of the reference model is refined horizontally (element side length from 25 m to 250 m) and vertically (layer thickness from 1 m to 30 m). The total number of nodes is 153,027.

Figure 3 A. As spatial discretization gets coarser, discharge simulation and, to a lesser extent, hydraulic head simulation is progressively deteriorated. **B.** While each model underestimates discharge during low flow periods, discharge during high flow periods is only overestimated by the coarsest models. This is highlighted by the graphs of weighted residuals.

Figure 4 A. As horizontal spatial discretisation gets coarser, NSE_q values are in general lower and RMS_h values are in general higher, this indicates that the simulation of both discharge and hydraulic heads are progressively deteriorated. **B.** The higher values of NSE_q and the lower values of RMS_h obtained with the calibrated models indicate that calibration significantly improves the simulation of both discharge and hydraulic heads.

Figure 5 Whatever the spatial discretization, the ranking of the most sensitive parameters and the magnitude of the composite scaled sensitivities are almost similar. This suggests that parameter sensitivities are not highly dependent on the grid size.

Figure 6 Calibration significantly improves model performance even for the coarsest models. **B.** Weighted residuals obtained with the calibrated models are almost randomly distributed. This indicates that calibration reduces model bias.

Figure 7 Values of NSE_q and RMS_h calculated for the validation period indicate that calibration also improves model performance outside the calibration period.

Figure 8 The 95% linear confidence intervals calculated for discharge only increase significantly for the coarsest model.

Figures 9 to 11 The 95% linear confidence intervals calculated for hydraulic heads are quite similar for each model.

Table captions

Table 1 Comparison of the number of nodes, number of elements, and execution times of the *250 m*, *500 m*, *750 m*, and *1000 m* models. The gain in execution time is tremendous when element size is increased.

Table 2 Values of NSE_q, MBE_q, $PE_q^{yr\,1}$, and $PE_q^{yr\,2}$ calculated for the *250 m*, *500 m*, *750 m*, and *1000 m* models. When spatial discretisation gets coarser, the variance of the hydrograph is poorly simulated (Gupta's 2nd term).

Table 3 Values of MBE_h, $HHVE_h^{yr\,1}$, and $HHVE_h^{yr\,2}$ calculated for the *250 m*, *500 m*, *750 m*, and *1000 m* models.

Table 4 Values of NSE_q, MBE_q, $PE_q^{yr\,1}$, and $PE_q^{yr\,2}$ calculated for the calibrated *250 m*, *500 m*, *750 m*, and *1000 m* models. Values in green are improved with regards to the corresponding models before calibration. Values in red are deteriorated with regards to the corresponding forward models.

Table 5 Values of MBE_h, $HHVE_h^{yr\,1}$, and $HHVE_h^{yr\,2}$ calculated for the calibrated *250 m*, *500 m*, *750 m*, and *1000 m* models. Values in green are improved with regards to the corresponding models without calibration. Values in red are deteriorated with regards to the corresponding forward models.

Table 6 Comparison of the reference value of the most sensitive parameters and their value after calibration.

Figure_1

Click here to download high resolution image

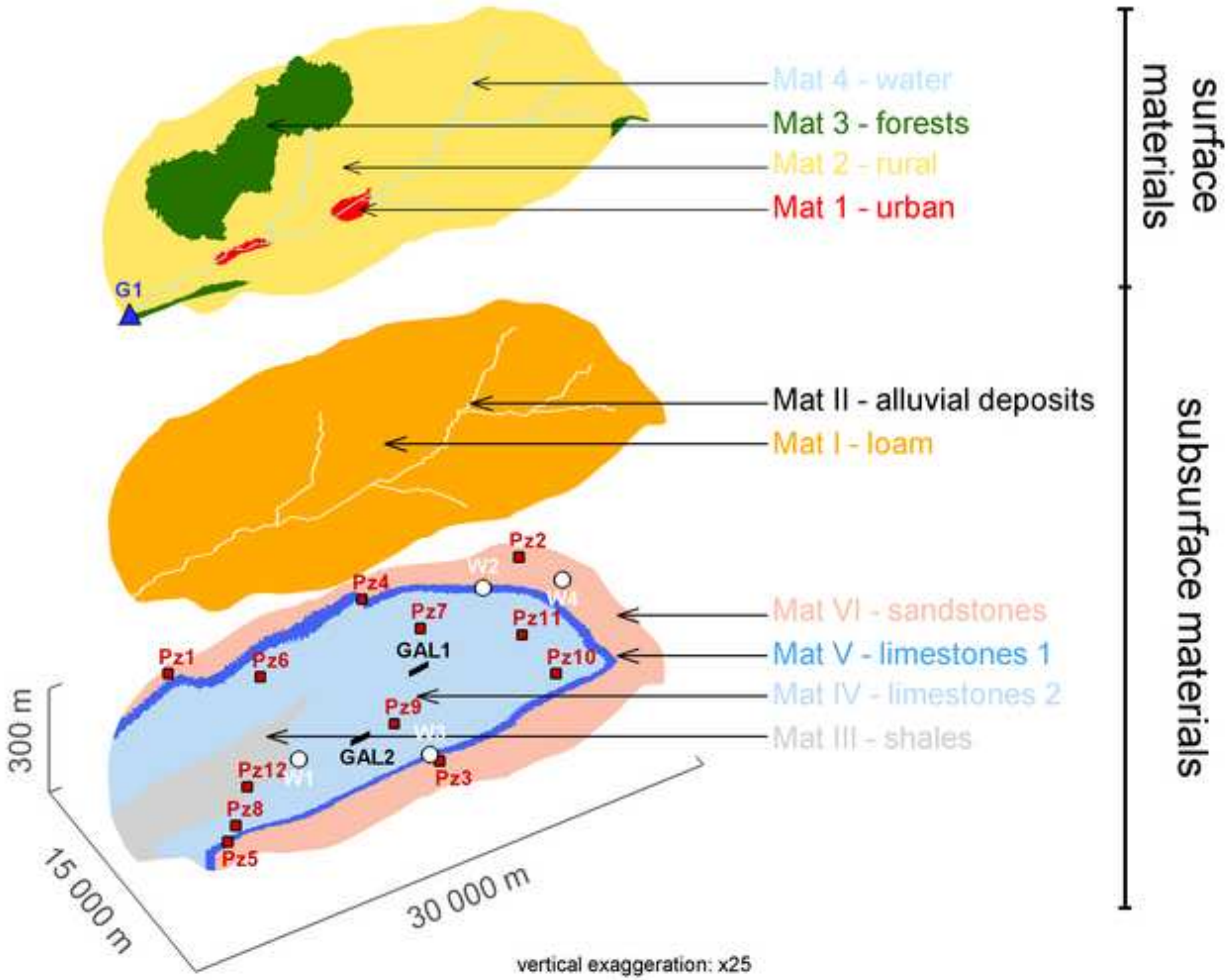

Figure_2
Click here to download high resolution image

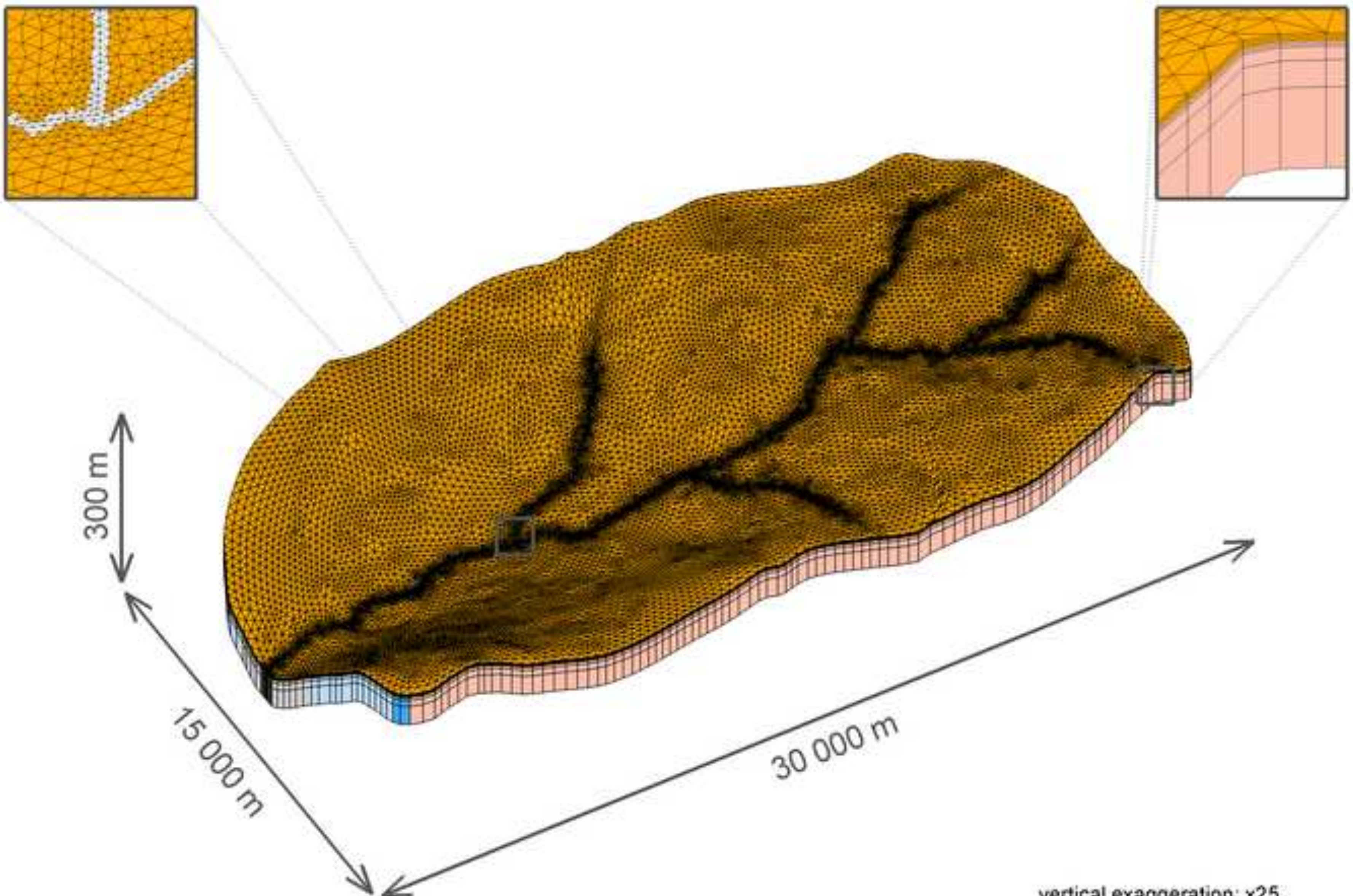

Figure_3

Click here to download high resolution image

Figure_4
Click here to download high resolution image

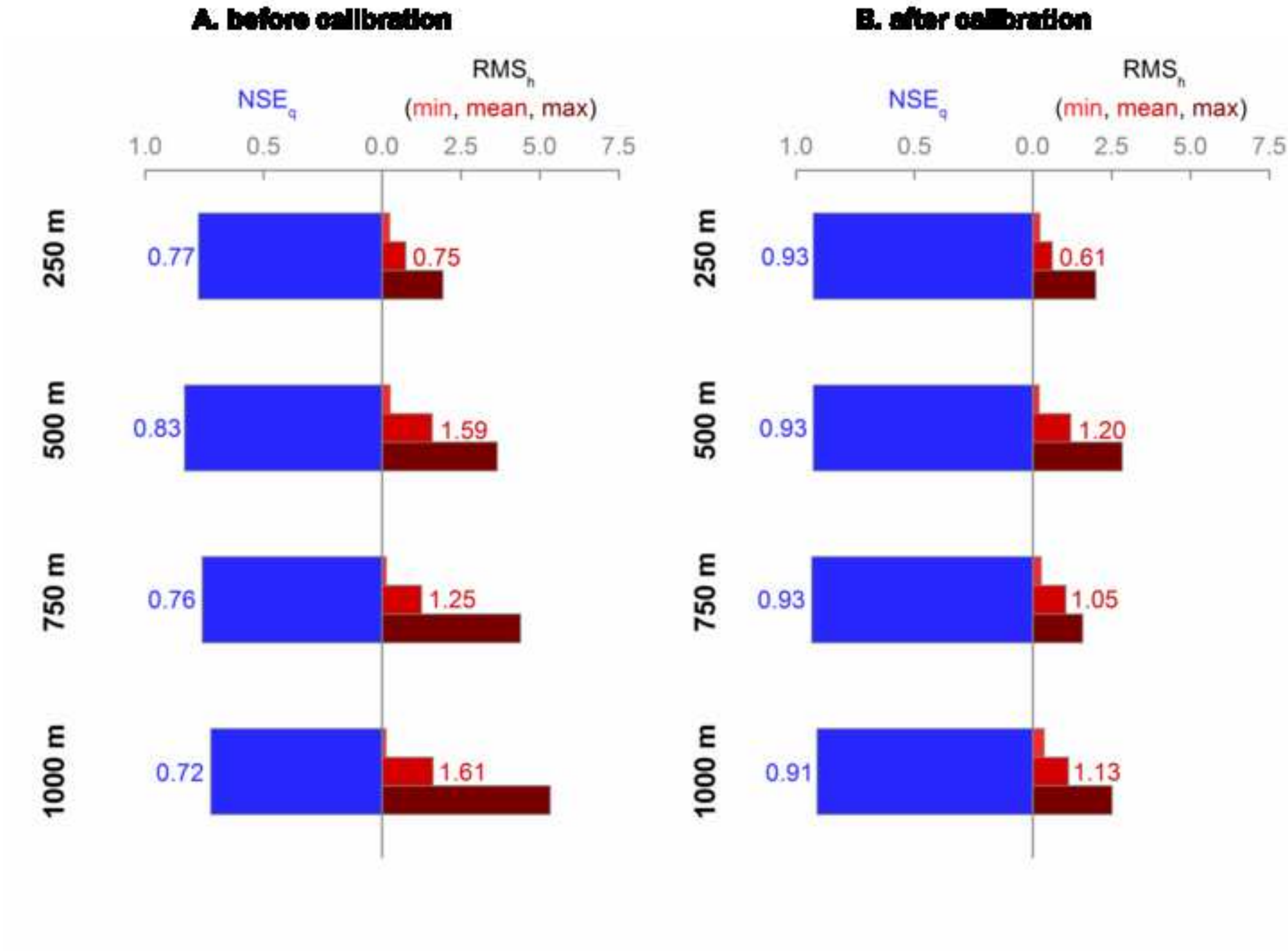

Figure_5
Click here to download high resolution image

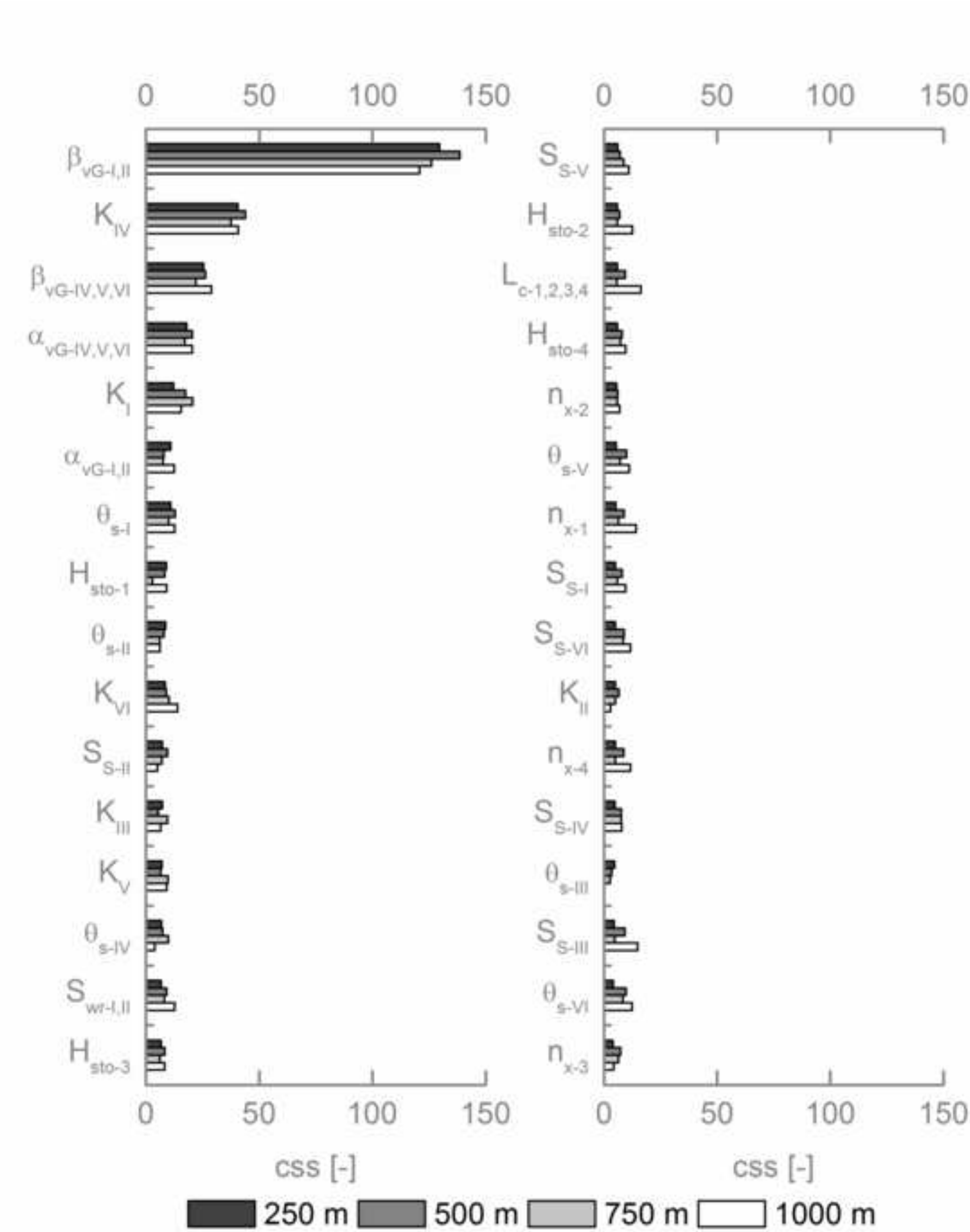

Figure_6
Click here to download high resolution image

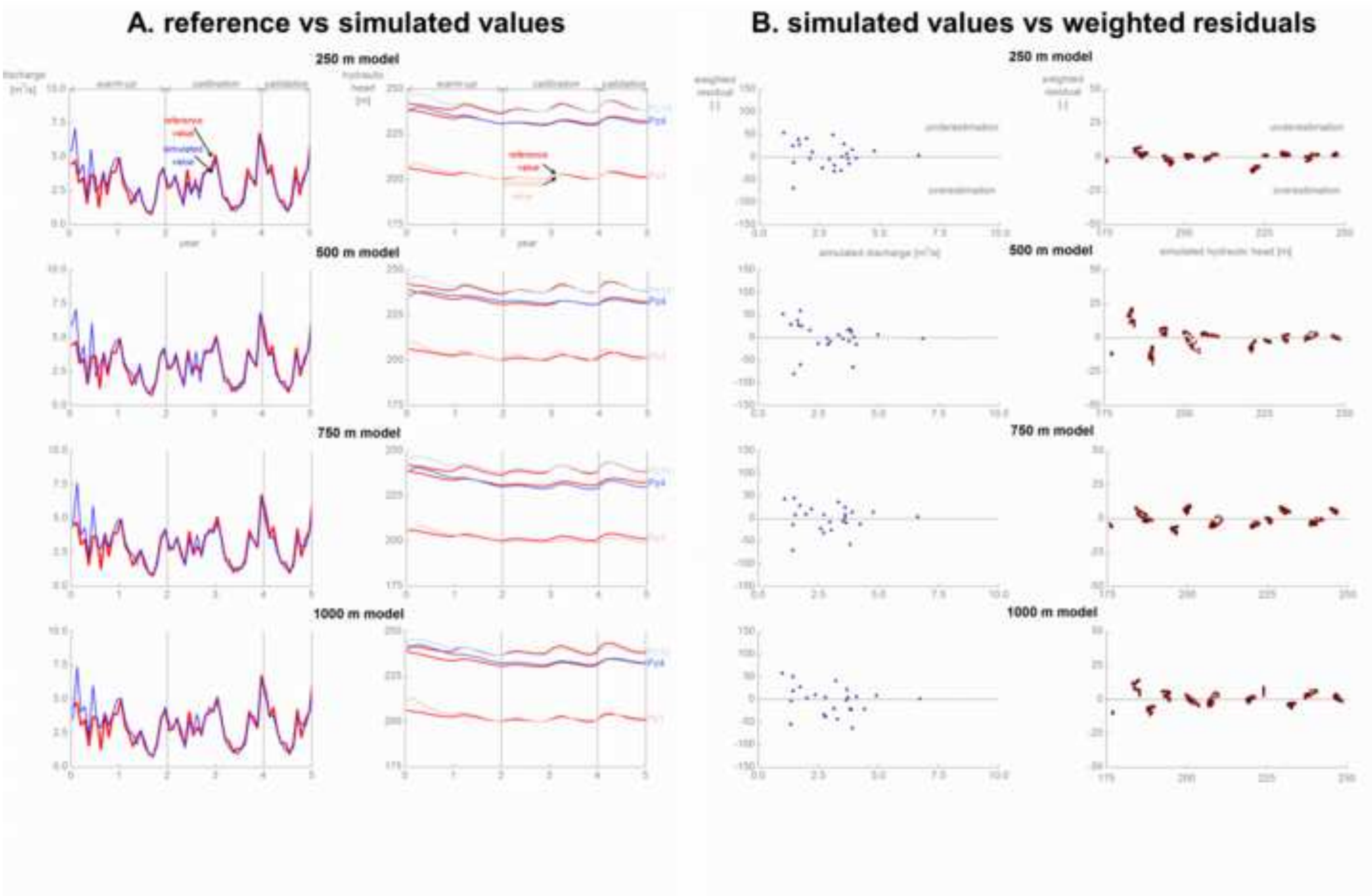

Figure_7
Click here to download high resolution image

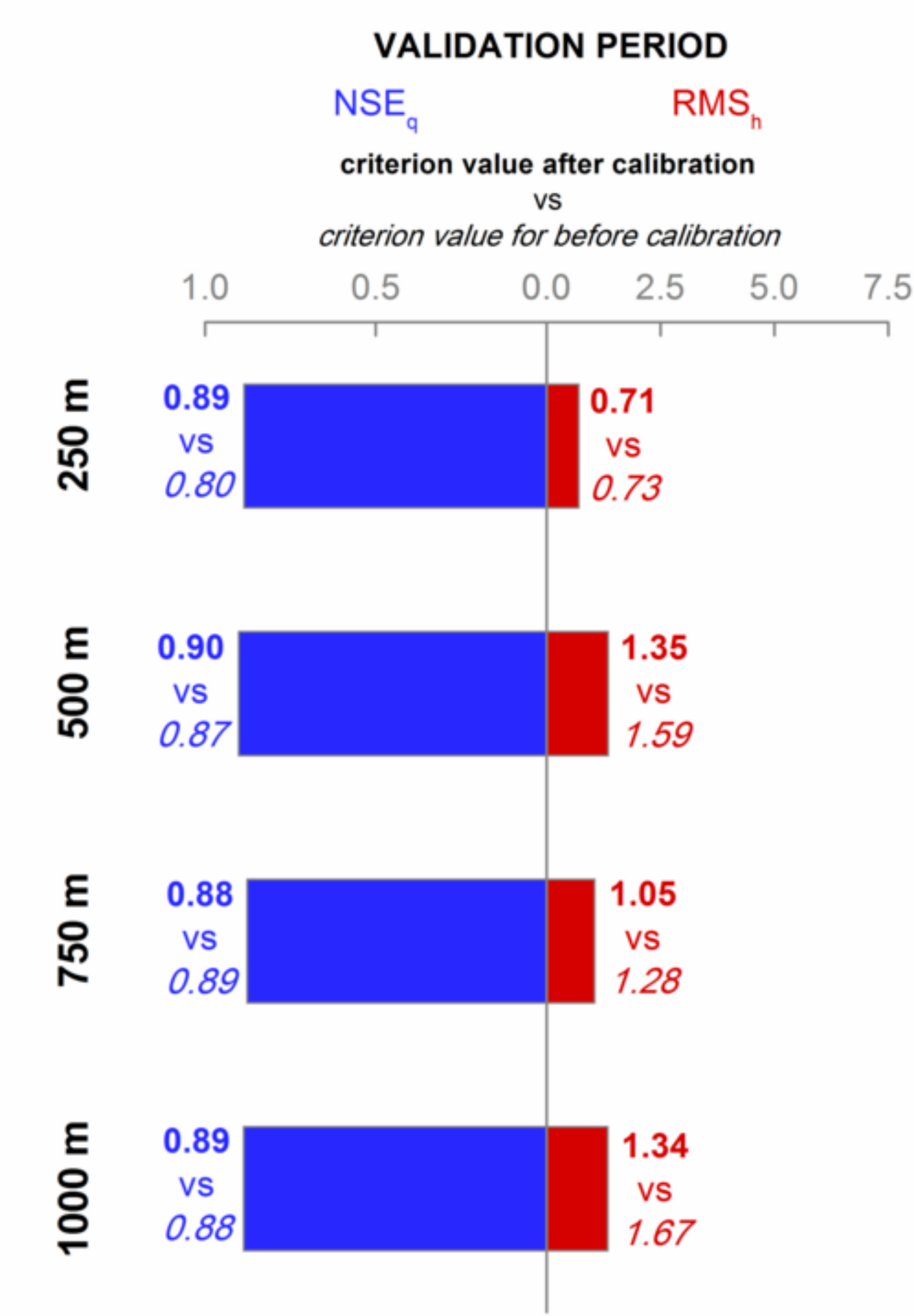

Figure_8
Click here to download high resolution image

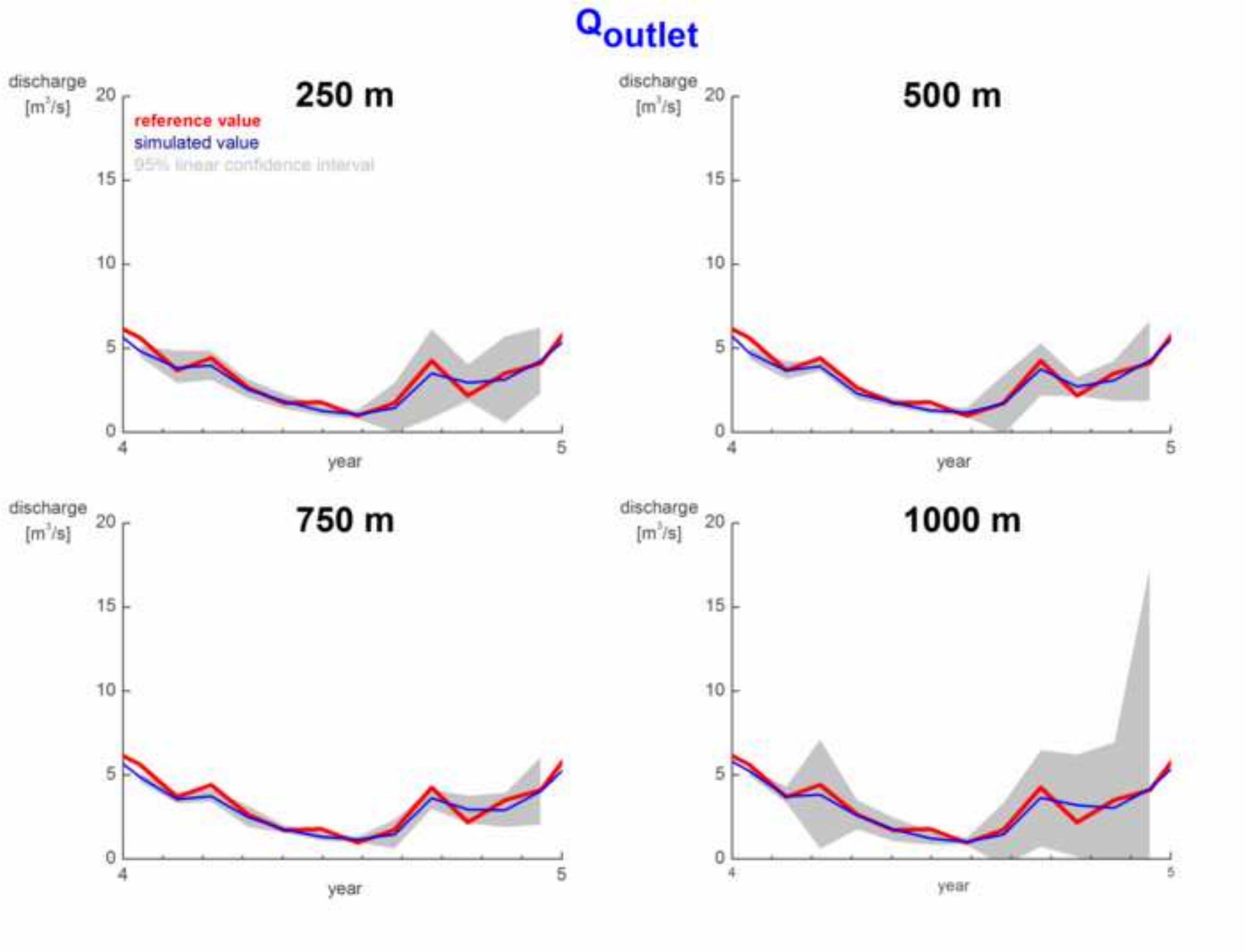

Figure_9
Click here to download high resolution image

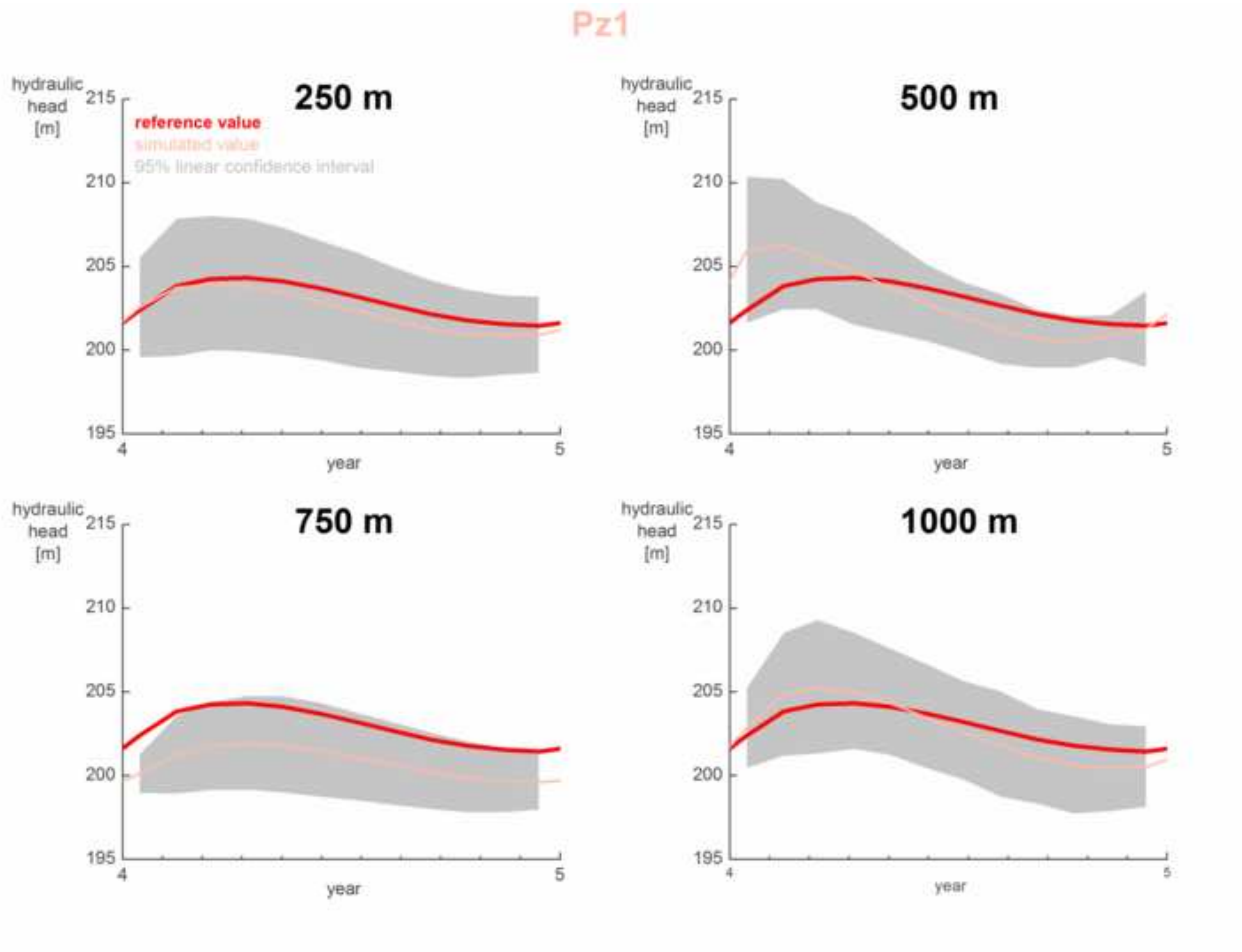

Figure_10
Click here to download high resolution image

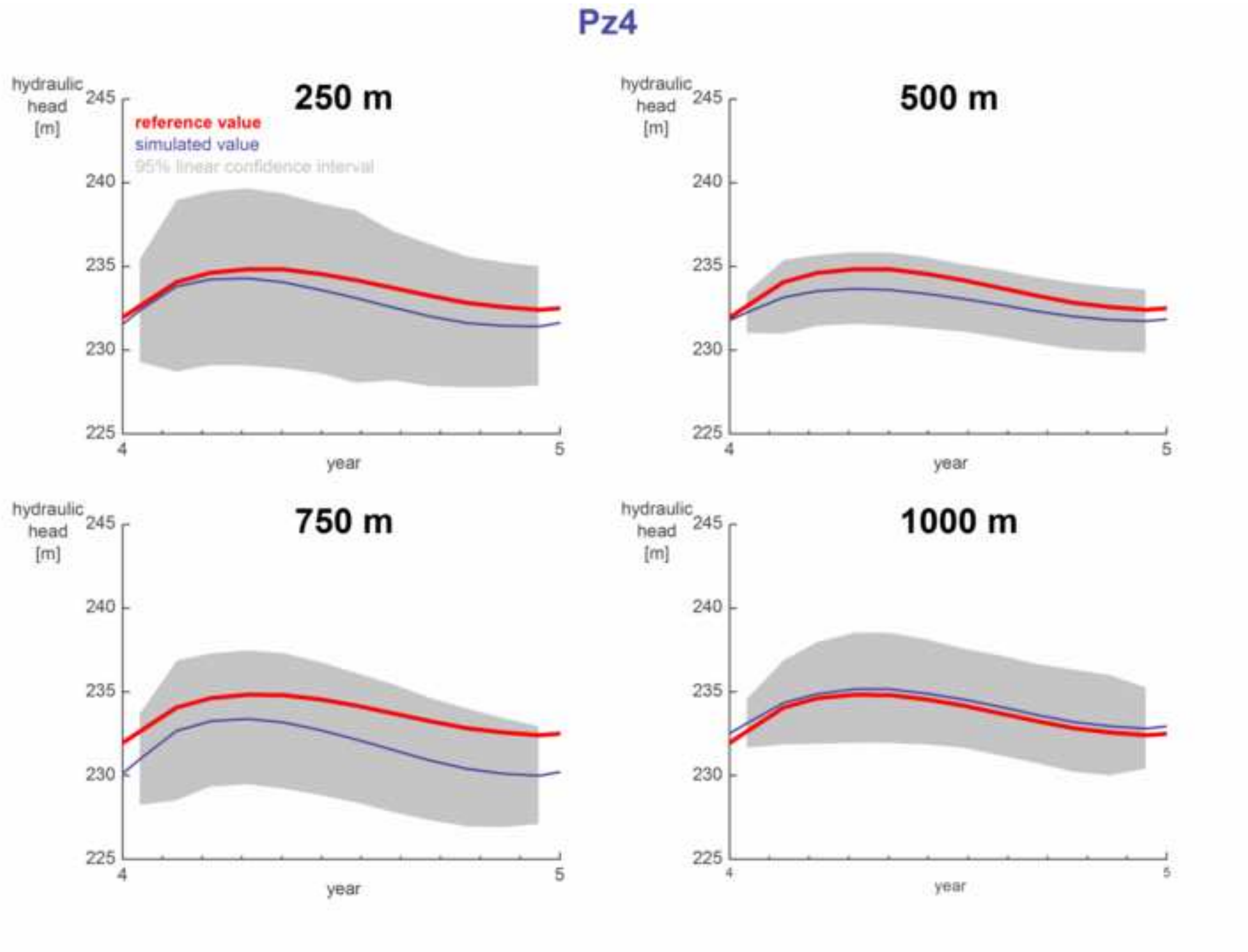

Figure_11

Click here to download high resolution image

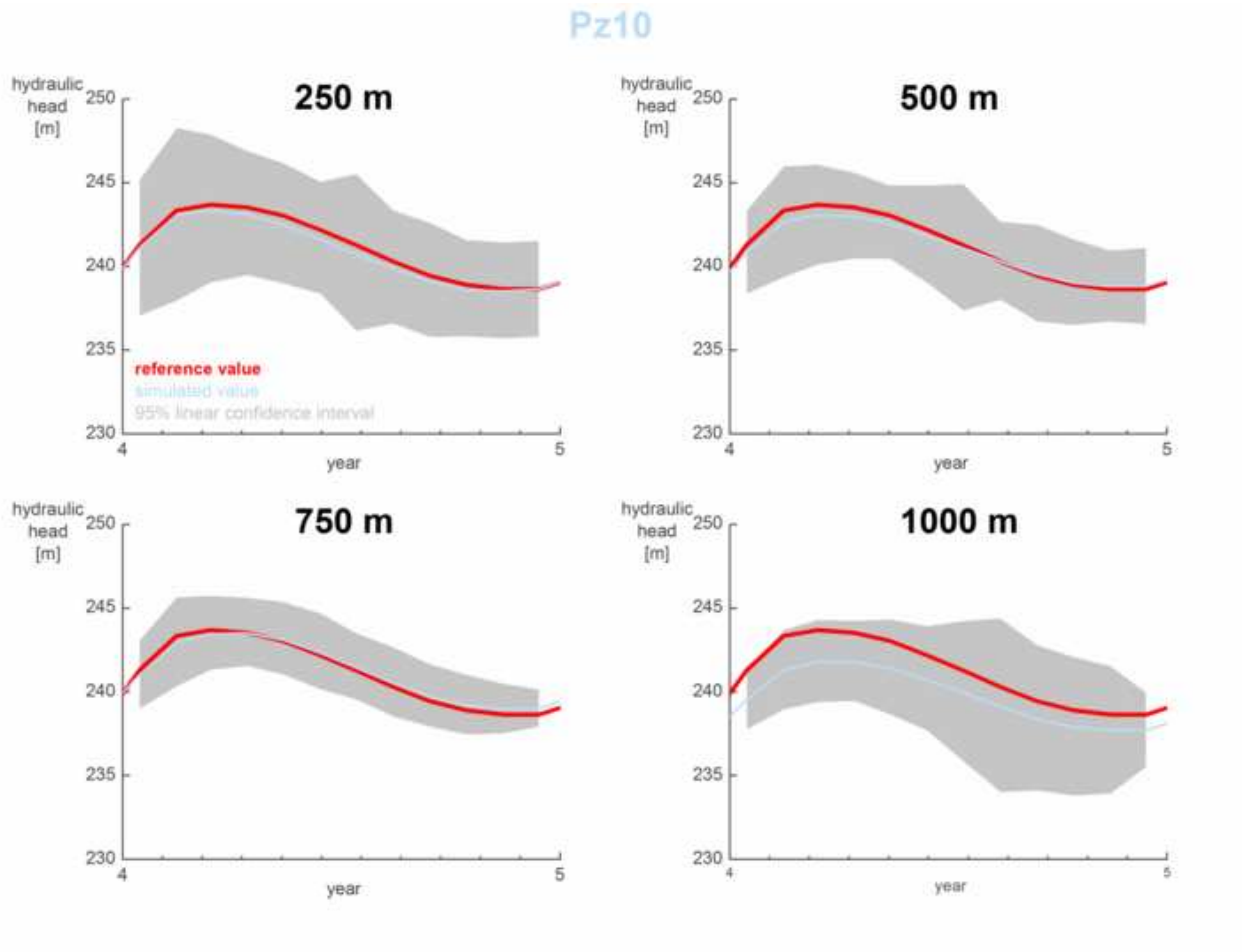

Table 1

	Number of nodes	Number of elements	Execution time
250 m	61,884	107,536	6.08 h
500 m	16,200	27,544	0.71 h
750 m	7,245	12,040	0.20 h
1000 m	4,302	7,016	0.13 h

Table 2

	250 m	*500 m*	*750 m*	*1000 m*
[-]	0.77	0.83	0.76	0.72
Gupta's 1st term [-]	2.26	2.33	2.53	2.59
Gupta's 2nd term [-]	1.37	1.47	1.77	1.87
Gupta's 3rd term [-]	0.11	0.03	0.00	0.00
[%]	-14.32	-7.00	-1.32	1.23
[%]	-8.65	-2.01	8.27	11.66
[%]	5.04	11.82	20.27	21.83

Table 3

		250 m	
	min	mean	max
[%]	-1.03	-0.17	0.50
[%]	-91.67	-9.40	24.39
[%]	-89.96	-17.69	3.66
		500 m	
	min	mean	max
[%]	-1.96	-0.15	1.18
[%]	-91.67	-4.22	51.35
[%]	-89.96	-12.62	24.70
		750 m	
	min	mean	max
[%]	-2.18	-0.36	0.49
[%]	-91.67	-12.46	19.51
[%]	-78.26	-19.83	2.76
		1000 m	
	min	mean	Max
[%]	-2.86	-0.62	0.96
[%]	-98.04	-21.46	21.62
[%]	-94.41	-24.97	10.98

Table 4

	250 m	*500 m*	*750 m*	*1000 m*
[-]	0.93	0.93	0.93	0.91
Gupta's 1st term [-]	1.87	2.01	1.87	1.91
Gupta's 2nd term [-]	0.94	1.08	0.94	1.00
Gupta's 3rd term [-]	0.01	0.00	0.00	0.00
[%]	-3.44	-0.28	-1.82	0.26
[%]	-4.81	-2.49	0.28	4.62
[%]	-2.44	1.71	-2.42	-4.25

Table 5

	min	mean	max
		250 m	
[%]	-0.48	0.06	0.90
[%]	-95.24	-15.08	49.62
[%]	-88.46	-7.57	38.43
		500 m	
	min	mean	max
[%]	-1.09	0.31	1.47
[%]	-85.71	-7.83	98.20
[%]	-80.77	-3.78	93.93
		750 m	
	min	mean	max
[%]	-0.96	-0.04	0.79
[%]	-36.11	5.14	52.74
[%]	-28.45	18.92	96.15
		1000 m	
	min	mean	max
[%]	-1.72	-0.16	0.93
[%]	-98.04	-21.40	61.26
[%]	-96.50	-16.89	63.16

Table 6

	reference model	***250 m***	***500 m***	***750 m***	***1000 m***
K_I [m/s]	5.00×10^{-7}	1.00×10^{-6}	1.66×10^{-6}	1.92×10^{-7}	7.38×10^{-7}
K_{IV} [m/s]	1.00×10^{-4}	9.91×10^{-5}	1.04×10^{-4}	9.75×10^{-5}	8.05×10^{-5}
$\alpha_{vG\text{-}IV,V,VI}$ [1/m]	3.65×10^{-2}	4.98×10^{-2}	4.56×10^{-2}	5.04×10^{-2}	3.04×10^{-2}
$\beta_{vG\text{-}I,II}$ [-]	1.45	1.47	1.32	1.53	1.39
$\beta_{vG\text{-}IV,V,VI}$ [-]	1.83	2.21	2.42	2.18	1.77