Advances in Water Resources 100 (2017) 109–125

Contents lists available at ScienceDirect

Advances in Water Resources

journal homepage: www.elsevier.com/locate/advwatres

Methods to correct and compute confidence and prediction intervals of models neglecting sub-parameterization heterogeneity – From the ideal toward practice

Steen Christensen

Department of Geoscience, Aarhus University, Høegh-Guldbergs Gade 2, 8000 Aarhus C, Denmark

ARTICLE INFO

Article history:
Received 15 April 2015
Revised 2 December 2016
Accepted 10 December 2016
Available online 12 December 2016

Keywords:
Heterogeneous media
Regression
Confidence interval
Prediction interval
Correction factor
Estimating weights during the regression

ABSTRACT

This paper derives and tests methods to correct regression-based confidence and prediction intervals for groundwater models that neglect sub-parameterization heterogeneity within the hydraulic property fields of the groundwater system. Several levels of knowledge and uncertainty about the system are considered. It is shown by a two-dimensional groundwater flow example that when reliable probabilistic models are available for the property fields, the corrected confidence and prediction intervals are nearly accurate; when the probabilistic models must be suggested from subjective judgment, the corrected confidence intervals are likely to be much more accurate than their uncorrected counterparts; when no probabilistic information is available then conservative bound values can be used to correct the intervals but they are likely to be very wide. The paper also shows how confidence and prediction intervals can be computed and corrected when the weights applied to the data are estimated as part of the regression. It is demonstrated that in this case it cannot be guaranteed that applying the conservative bound values will lead to conservative confidence and prediction intervals. Finally, it is demonstrated by the two-dimensional flow example that the accuracy of the corrected confidence and prediction intervals deteriorates for very large covariance of the log-transmissivity field, and particularly when the weight matrix differs from the inverse total error covariance matrix. It is argued that such deterioration is less likely to happen for three-dimensional groundwater flow systems.

1. Introduction

A groundwater system can have complicated structure and possess heterogeneity within its structural elements (termed structures in the following). When the structures are known a model can be built to simulate groundwater flow provided that boundary conditions, sources, sinks, and hydraulic properties within the structures can be specified. Because of data scarcity the spatial distribution of hydraulic properties within the structures will always be unknown to some (usually large) degree. This introduces uncertainty in predictions made by groundwater model simulation. One possible way to quantify prediction uncertainty is by high-resolution Monte Carlo (MC) simulation, which applies directly to nonlinear problems by sampling from the assumed known statistical distribution of the hydraulic property fields. However, MC simulation is computationally expensive when the simulations also need to be constrained by, for example, hydrological data observed in the field (Neuman, 2003). The variant called Markov chain Monte Carlo (McMC) simulation is also expensive (see for example demonstration by Lu et al., 2012), and it can suffer from imprecision caused by inadequate sampling (Lu et al., 2012). As argued and stated by Cooley and Christensen (2006), and supported by Lu et al. (2012), the Monte Carlo method is therefore no panacea in groundwater modeling.

Alternatively the model can be simplified by ignoring (all or some) heterogeneity within the structures, so only a relatively small number of property trend values need to be specified or estimated – e.g. the average transmissivity within each structural element, or the transmissivity at a small number of pilot points within each structural element. Some or all of these property trend values are often estimated by nonlinear regression using hydrological field observations as calibration data. Ignoring variability around the trend introduces the kind of model error studied by Cooley (2004), Cooley and Christensen (2006), and here. Cooley (2004) and Cooley and Christensen (2006) showed that such model error increases the variance of the total errors (where total error is the sum of model error and observation error) as well as the correlation between the total errors of simulated equivalents to observations and model predictions. It also causes bias in the

E-mail address: sc@geo.au.dk

http://dx.doi.org/10.1016/j.advwatres.2016.12.007
0309-1708/

regression-estimated trend parameters, but usually not in simulations of parameter dependent variables and predictions (see sections on intrinsic nonlinearity in Cooley, 2004, and Cooley and Christensen, 2006). Finally it can make traditionally calculated, here termed *uncorrected*, confidence and prediction intervals (e.g. Graybill, 1976; Vecchia and Cooley, 1987; Hill and Tiedeman, 2007) very inaccurate when the statistical distribution of the total errors is different from that implied by the weight matrix used for regression and for calculation of the intervals.

Cooley (2004) and Cooley and Christensen (2006) also show that if the within-structure heterogeneity that is neglected in the model can be treated as a random field with known geostatistical properties then unconstrained high-resolution MC simulation (which is inexpensive compared to calibration-constrained MC simulation) can be used to estimate statistical moments for the total errors which makes it possible to correct and thereby improve the accuracy of confidence and prediction intervals. Such from now on termed *corrected* intervals quantify uncertainty in parameters or predictions due to both observation error and model error caused by neglecting heterogeneity. So the capability of this regression-based method to quantify uncertainty of model predictions will often be comparable to the capability of the calibration-constrained MC method, but the computational expense of the former method will generally be much less than for the latter. This was demonstrated in the study of Christensen et al. (2006) who used both the regression-based method and a calibration-constrained MC method (Doherty, 2003) to quantify uncertainty of a prediction made by a groundwater model of a synthetic heterogeneous aquifer. Similarly, Lu et al. (2012) showed that regression based confidence intervals and McMC based credible intervals (also termed "credibility intervals") will often be similar but the computational expense to compute the latter exceeds the expense to compute the former by, in their examples, three orders of magnitude. Furthermore, Lu et al. (2012) also argue that "it may be useful to calculate the less computationally demanding confidence intervals early in the development, and calculate the computationally demanding credible intervals as the model becomes a better representation of the system".

Confidence and prediction intervals can also be computed by using the Monte Carlo and regression based percentile bootstrap method (Efron, 1982; Stine, 1985; Cooley, 1997). However, the computational expense of computing such bootstrap intervals will be orders of magnitude larger than of computing the corrected regression based intervals mentioned above (Cooley, 1997). The reason is that regression must be repeated for each of many (possibly thousands of) bootstrap samples. Furthermore, bootstrap intervals may be inaccurate when the bootstrap distribution is not a translation of the true distribution; see Cooley (1997) for a groundwater modeling example.

It is typical for many (or most) field cases that the data on hydraulic properties are few and clustered. The geostatistical properties of heterogeneous hydraulic property fields are therefore difficult or impossible to estimate from field data for all relevant spatial scales. In other words, it is rarely possible to postulate what Neuman (2003) calls a Type A probabilistic model of prior hydraulic property uncertainty which requires an extensive set of local field hard data. Alternatively subjective judgment must be used together with indirect information (e.g. parameter measurements from similar hydrogeologic environments and/or geophysical data indicative of hydraulic properties) to suggest a Type B probabilistic model of prior property uncertainty (Neuman, 2003). Type B model geostatistical properties are prone to be biased and have wrong variance due to lack of local data and/or because geophysical data may be poorly correlated to, or of other scale than, the relevant hydraulic property values. In almost any real case the proposed probabilistic model for the hydraulic properties will to some extent be of either Type B or intermediate between Type A and Type B. In practice this leaves the MC and the regression based methods to always be approximate methods of quantifying uncertainty of model parameters and predictions. Because the geostatistical properties are at least partly unknown an obvious alternative to the regression-based and the MC methods is the quasi-linear geostatistical method (Kitanidis, 1995) because this estimates the heterogeneous hydraulic property field as well as parameters of the property field covariance function. However uncertainty must be quantified by MC simulation using the randomized maximum likelihood method (Oliver et al., 2008, p. 320–334) which is computationally very expensive. Furthermore, for high-resolution modeling it requires a very efficient procedure for computation of the sensitivity (Jacobian) matrix and for computing and storing huge covariance matrices (Nowak and Cirpka, 2004); particularly the latter may be prohibitive for using the geostatistical approach unless approximate methods (e.g. those of Kitanidis and Lee, 2014) are brought into play. In most cases the regression-based method will therefore probably be the most efficient computational method for quantifying uncertainty in relation to high-resolution groundwater modeling. Its main drawback is that high intrinsic nonlinearity causes problems with interpreting parameter values and with accuracy of uncertainty measures (Cooley, 2004; Cooley and Christensen, 2006), but it is suspected that any method would be similarly affected. For the MC method, for example, both the problem of adequate sampling and the model runtime are likely to increase with the nonlinearity of the problem.

The total error variance for observations used for parameter estimation can be estimated from field data, field evidence of large-scale heterogeneity, grouping of data, and residual analysis (see e.g. Christensen, 1997; Christensen et al., 1998; Christensen and Cooley, 1999; Cooley and Naff, 1990). For two field cases Christensen and Cooley (1999) used such estimated total error variances to form the diagonal weight matrix used for parameter estimation by regression and to calculate uncorrected prediction intervals for predictions of same types as data used for the estimation. They tested and could not reject the hypothesis that the uncorrected prediction intervals are accurate. The theory presented by Cooley (2004) and Cooley and Christensen (2006) supports the idea that prediction intervals often need relatively little correction to become accurate whereas confidence intervals can be much too small unless they are corrected by a correction factor.

In some studies the weights are computed as part of the regression (e.g. Wagner and Gorelick, 1986, 1987; Barlebo et al., 1998). This conforms to the assumption that the total error variances are unknown but the residuals contain information about their values. However, because residuals depend on the weights used, the weights should be estimated simultaneously with the model parameters in order for the weights to best estimate the inverse of total error variances. This was formally shown by Cooley (2004, p. 40–41). He also developed and tested procedures for computing approximate confidence and prediction intervals in this instance (Cooley, 2004, p. 55–56, p. 67–68, and p. 109–115). Confidence and prediction intervals also in this case need correction to account for uncertainty due to model error.

This paper continues the work presented by Cooley (2004) and Cooley and Christensen (2006). In their work it is assumed that Type A probabilistic models can be postulated for the heterogeneous fields forming the groundwater system. Here this assumption is step-wise relaxed and the calculation and correction of confidence and prediction intervals are modified accordingly. It is shown how the traditional calculation of confidence and prediction intervals can be corrected in alternative ways. Some correction alternatives correspond to those developed by Cooley (2004), others are new. The results are subsequently rewritten and simplified in two steps. In step 1 they are simplified to allow making corrections when the total error variances can be estimated as