

S.S. Papadopoulos & Associates, Inc. Short course notes:
Practical Modeling of Pump-and-Treat Systems Using MODFLOW, PATH3D and
MT3D

Notes on the calibration of groundwater models

Christopher J. Neville
S.S. Papadopoulos & Associates, Inc.
October 1994

Additional material on the Spitz and Moreno model calibration acceptance criterion
appended May 27, 2008

Overview:

These notes have been prepared to provide guidance regarding model calibration and may be a resource during the initial stages of model development, during the development of a formal calibration strategy, and for the assessment of the goodness-of-fit of a model.

The notes are divided into five main sections:

1. Definition of model calibration;
2. Calibration philosophy and strategy;
3. Calibration targets;
4. Evaluation of model goodness-of-fit; and
5. Other considerations in assessing whether a model is fit for purpose.

1. Definition of model calibration

Model calibration is the process of adjusting selected model structure and parameters systematically to match observed data. The parameters are adjusted systematically within physical bounds. Calibration may also require adjustment of the model structure.

In an ideal world, a groundwater simulation would replicate exactly the groundwater system with model parameters specified completely from independent evaluations. In such a world, we would not require a discussion of the adjustment of model parameters or an evaluation of the match between observations and model calculations. In reality, the natural groundwater system is dynamic and can never be characterized completely. The measurements we make can never be truly representative. Water levels in wells may change rapidly, and errors may arise in their measurement. Furthermore, a well is an imperfect instrument for determining water levels in the subsurface. Parameter estimates are generally sparse, and determined from tests that “sample” the properties of the subsurface at different scales. The interpretations of these tests are generally based on idealized analyses, and are inherently approximate.

The confidence with which we can use a numerical groundwater model for predictive purposes is strengthened if it can be shown that the model can match observed water levels for different conditions, and the model parameters are constrained by the available data. Since a groundwater model can provide only an approximate representation of a complex physical system, calibration of the model and an evaluation of the match between the observations and the model calculations is an essential part of model development.

2. Calibration philosophy and strategy

The most defensible model is obtained when as few parameters as possible are adjusted and all of the available data are considered.

The process of model calibration should be approached with a deliberate reluctance to add complexity to the model unless clearly warranted. A key guideline has been to make as many simplifications as necessary, but not more. We believe that this is the single most important element in developing a model that is useful for predictive purposes. A model that incorporates all available detail is generally unwieldy and may ultimately be useless.

Three key questions that must be addressed in developing a calibration strategy are:

1. What are the calibration targets?
2. What parameters will be adjusted? and
3. How will the parameters be adjusted?

With regards to deciding what parameters will be adjusted, we should establish:

- What parameters must be specified?
- What parameters will be assigned fixed values based on independent estimates?
- What parameters can be assigned fixed values based on literature ranges (i.e., assuming reasonable “textbook” values)?
- Of the parameters that will be adjusted, what information is available that can constrain their values? Can any of this information be used to establish ranges over which parameter values may be adjusted?
- How will the values of the distributed parameters be zoned?

Model calibration should be approached through a “trial-and-study” approach. Results of alternative model structures are carefully reviewed to evaluate the consistency of the model results with the conceptual model. For each alternative structure investigated, the final model calibration generally requires a combination of manual parameter adjustments, and the application of a model-independent, non-linear parameter estimation code, such as PEST (Doherty, 2004).

3. Calibration targets

In a typical groundwater flow model, the objective of calibration is to adjust the model so that it achieves an acceptable match to observed water levels. However, many observations may serve as appropriate targets for the calibration of groundwater models, including water levels, discharge rates, groundwater age, and solute concentrations and travel times.

The agreement between a model and an individual observation (target) is expressed as a model residual, defined as the difference between the calculated and observed value:

$$R_i = cal_i - obs_i \quad (1)$$

According to the sign convention of the definition above, a positive residual indicates over-calculation by the model, and a negative residual indicates a model under-calculation. Not all modeling packages adopt this sign convention. For example, some programs define it as the difference between observed and calculated water levels. It is important to understand what positive and negative residuals represent in a particular context.

In a particular context, it may also be appropriate to assign different weights to the residuals. This will have the effect of placing more emphasis on achieving a match to some targets and discounting the importance of other targets. For example, more emphasis might be directed towards matching water levels derived from time series from dedicated monitoring wells than spot measurements from private water wells at the time they were drilled. Hill (1998) defines the weighted residual as:

$$R_i = (cal_i - obs_i) w_i^{1/2} \quad (2)$$

Here w is the weight assigned to observation i . The weights may serve more than one purpose:

- The weights may account for the reliability we attach to specific observations in the calibration; and/or
- The weights may be used to allow combined consideration of different types of measurements that have different units and significantly different magnitudes (heads and flow rates for example).

4. Evaluation of model goodness-of-fit

How do we know when we can stop adjusting the model parameters? In other words, how do we know when the match between the observations and model results is acceptable? There is no standard methodology for making such an assessment. The calibration process necessarily requires the judgement of an expert modeler. Here we focus on the techniques that are used to compare the results of a model against observations. The American Society for Testing and Materials (ASTM) has published guidelines for comparing model results against observations (ASTM Standard Guide D5490-93). In our opinion, the ASTM guidelines are a good starting point, but should be used **only** as a starting point.

Model calibration is evaluated using both quantitative and qualitative measures. Quantitative evaluation generally consists of assessing the statistics of the residuals of a simulation. Qualitative evaluation consists of judging whether the results of the modeling are sensible and consistent with the overall conceptual model. Although qualitative measures necessarily involve judgement, they are not inferior to quantitative measures. Rather, they should be regarded as being complementary.

4.1 Quantitative measures of goodness-of-fit

Statistical measures calculated with the residuals provide the modeler with a means of determining whether revisions to the model have yielded a demonstrably better match to the observations. However, it is important to note that no single statistic adequately quantifies the match between observations and model calculations. Eight alternative cumulative measures of goodness-of-fit are assembled below. All of the measures may be expressed in terms of either the original or weighted residuals.

1. Mean residual (MR)

The mean residual is defined as the arithmetic average of the residuals:

$$ME = \frac{1}{N} \sum_{i=1}^N (cal_i - obs_i) \quad (3)$$

Here N is the number of observations; cal_i is the calculated value; and obs_i is the observed value.

Guidance:

It is desirable that the mean residual be relatively close to zero, although we have not found any definitive guidance on what is “close” in this context. A mean residual of zero indicates that the calculated water levels are at the right average level. Although a mean residual that is close to zero is desirable, it is not a sufficient indicator of a good match, because large negative and positive residuals may cancel out.

2. Mean absolute residual (MAE)

The mean absolute residual is defined as the arithmetic average of the absolute value of the residuals:

$$MAE = \frac{1}{N} \sum_{i=1}^N |cal_i - obs_i| \quad (4)$$

The mean absolute residual is a more useful measure of overall goodness-of-fit than the mean residual, as positive and negative residuals do not cancel. This is a measure in some sense of how wrong the model might be average at any one location in the model.

Guidance:

It is desirable that the absolute residual be relatively close to zero: A mean residual of zero indicates that the calculated water levels are at the right average level, and that the scatter of residuals is relatively small.

We have not found any guidance in the literature on what “close to zero” means.

3. Standard deviation of residuals (SDEV)

$$SDEV = \left[\frac{1}{N-1} \sum_{i=1}^N (cal_i - obs_i)^2 - \frac{N}{N-1} MRE^2 \right]^{1/2} \quad (5)$$

The standard deviation of the residuals provides a quantitative measure of the scatter of the residuals. As defined here, incorporating $N-1$ instead of N in the denominator, $SDEV$ represents the “unbiased” approach for estimating the standard deviation of a sample of the entire population.

Guidance:

It is desirable that the standard deviation be relatively close to zero. A standard deviation that is close to zero indicates that the calculated water levels are both at the right average level, and that the scatter of residuals is relatively small.

We have not found any guidance in the literature on what “close to zero” means.

4. Root-Mean-Squared Residual (RMSE)

$$RMSE = \left[\frac{1}{N} \sum_{i=1}^N R_i^2 \right]^{1/2} \quad (6)$$

The *RMSE* is closely related to the mean absolute residual and the standard deviation. The value of the *RMSE* is close the standard deviation when the mean error, *ME*, is small and the number of observations, *N*, is relatively large.

Guidance:

As with the standard deviation of the residuals, it is desirable that the RMSE be relatively close to zero. An RMSE deviation that is close to zero indicates that the calculated water levels are both at the right average level, and that the scatter of residuals is relatively small.

We have not found any guidance in the literature on what “close to zero” means.

5. Normalized root-mean-squared error (nRMSE)

There is no fixed definition of the normalized root mean squared error. One definition of the normalized RMSE is:

$$NRMSE = \frac{1}{\bar{O}} \left[\frac{1}{N} \sum_{i=1}^N R_i^2 \right]^{1/2} \quad (7)$$

The term $\bar{O}$ represents the *average* of the observed values $\bar{O} = \frac{1}{N} \sum_{i=1}^N O_i$

Another definition of the normalized RMSE is:

$$NRMSE = \frac{1}{(O_{\max} - O_{\min})} \left[\frac{1}{N} \sum_{i=1}^N (C_i - O_i)^2 \right]^{1/2} \quad (8)$$

The term $(O_{\max} - O_{\min})$ represents the *range* in the observed heads.

Guidance:

Spitz and Moreno (1996; corrected 2003) suggest that the $nRMSE$ should be less than 10% of the range of the observations.

$$\frac{SDEV}{obs_{\max} - obs_{\min}} < 0.1 \quad (9)$$

See additional discussion on the correct interpretation of the Spitz and Moreno “model acceptance criterion” appended to these notes.

6. Sum of squared residuals (SSE)

The sum of the squared errors, SSE , (more precisely the sum of the squared weighted residuals) is defined as:

$$SSE = \sum_{i=1}^N (cal_i - obs_i)^2 \quad (10)$$

Here N is the number of observations.

The magnitude of the sum of the squared residuals depends on the number of observations. Therefore, it is not possible to provide guidance with respect to its value for model acceptance. However, it is important to note that the sum of the squared errors is still a crucial quantity. The objective of computer-assisted methods is to minimize the sum of the squared residuals, or the sum of the weighted squared residuals. In other words, the match between a model and the observations cannot be improved with further adjustment of parameter values when no change in the sum of the square residuals is achieved between successive simulations.

7. Pearson linear correlation coefficient (r)

$$r = \frac{\sum_{i=1}^n (obs_i - \bar{o})(cal_i - \bar{c})}{\left[\sum_{i=1}^n (obs_i - \bar{o})^2 \right]^{1/2} \left[\sum_{i=1}^n (cal_i - \bar{c})^2 \right]^{1/2}} \quad (11)$$

Zheng and Bennett (2002; p. 320, Equation 12.5) designate this statistic as the linear correlation coefficient.

Guidance:

Spitz and Moreno (1996; corrected 2003) have suggested that the correlation coefficient should lie between 0.7 and 1.0.

$$r = 0.7 - 1.0 \quad (12)$$

A closely related measure is the Coefficient of Determination (CD):

$$CD = \frac{\frac{1}{N} \sum_{i=1}^N (O_i - \bar{O})^2}{\frac{1}{N} \sum_{i=1}^N (C_i - \bar{C})^2} \quad (13)$$

The term $\bar{C}$ represents the average of the observed values. $\bar{C} = \frac{1}{N} \sum_{i=1}^N C_i$, and $\bar{O}$ is defined similarly.

We have not found any guidance in the literature on criterion for model acceptance based on the CD .

4.2 Qualitative measures of acceptable model calibration

Qualitative assessment is made with visual comparisons and may include examination of:

- Examination of the overall match between observed and calculated water levels;
- Examination of residuals at individual wells;
- Evaluation of the shape of the water level contours;
- Evaluation of flow patterns;
- Evaluation of hydraulic gradients;
- Evaluation of absolute groundwater levels; and
- Evaluation of consistency between the calibrated parameters and parameter estimates that have been obtained independently of the calibration.

Qualitative evaluation includes the visual comparison of groundwater flow directions and key hydrologic features, such as groundwater flow divides. Qualitative evaluation also includes checking the calculated flow balance to ensure that its components fall within physically realistic bounds that are consistent with the site conceptual model.

1. Scatterplots

Qualitative evaluation includes the preparation of scatterplots of computed versus observed water levels. Scatterplots are a standard method of providing a visual impression of the quality of fit for a steady-state model (American Society for Testing and Materials, 1993). In a standard scatterplot, the abscissa (x-axis) represents the observed water level, and the ordinate (y-axis) represents the calculated water level at the location of the well. An example scatterplot is shown in Figure 1. For a perfect match between the model calculations and the observation, the points fall on a straight line. The points on a scatterplot should be scattered randomly about the line of perfect agreement, with a relatively narrow scatter about that line.

The scatterplot is inspected for deviations of the data from the line of equality. The data points on the scatter plot are also inspected for patterns. For example, if the slope of a line through the data points is steeper or shallower than the solid line, this suggests an error in the overall hydraulic gradient. As with the visual comparison of water level contours, examination of the scatter plot is a qualitative assessment.

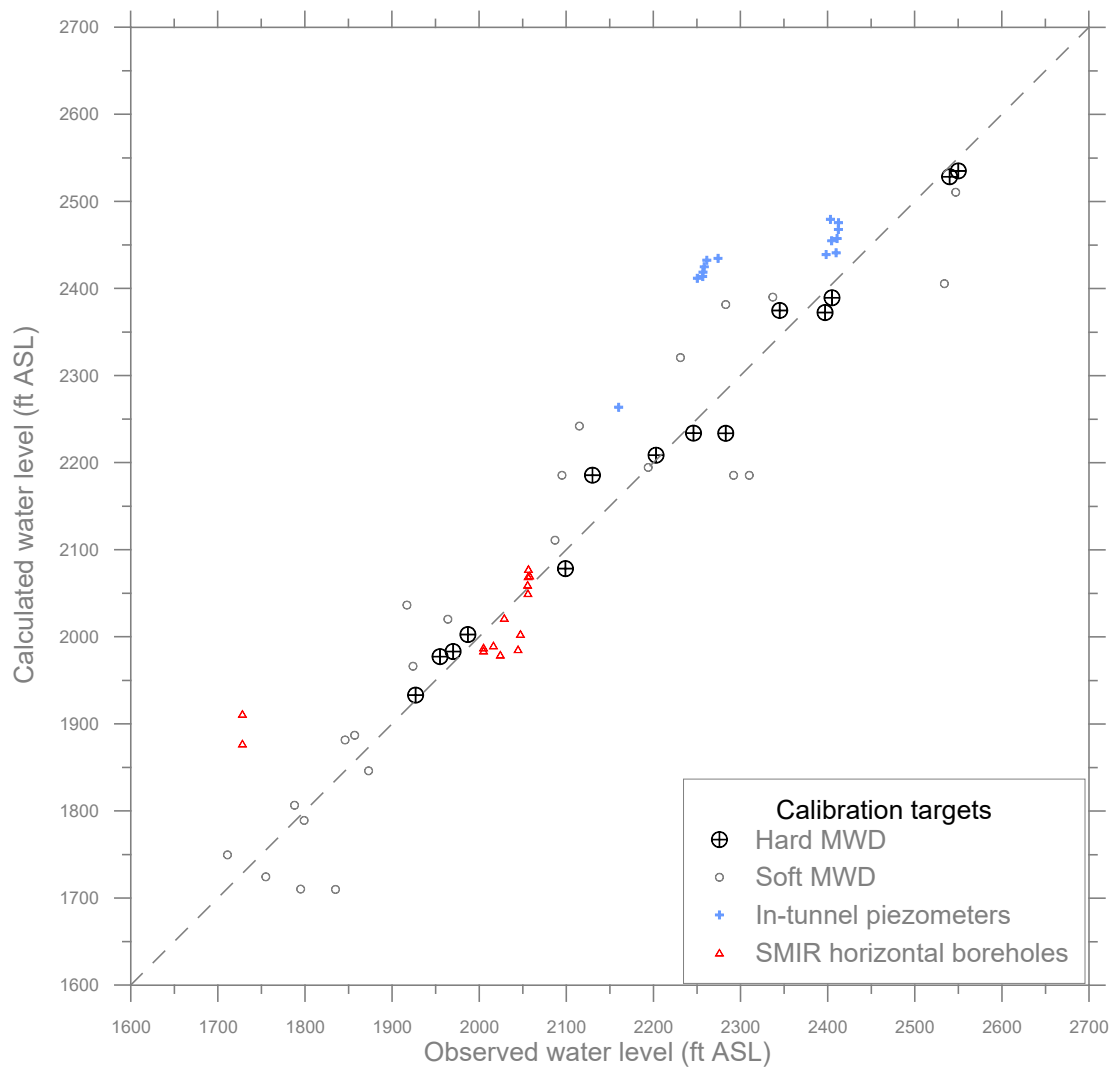


Figure 1. Example scatterplot of head residuals

2. Cumulative probability plot of model residuals

Spitz and Moreno (1996, p. 244-245) and Hill (1998) have suggested that the residuals from a calibration should be normally distributed, with a mean of zero. We interpret this to mean that the largest portion of the residuals plotted on a probability plot should approximate a straight line, with the residual corresponding to 50% close to zero. An example probability plot of model residuals is shown in Figure 2. The cumulative probability plot is useful for identifying targets that are likely unreliable.

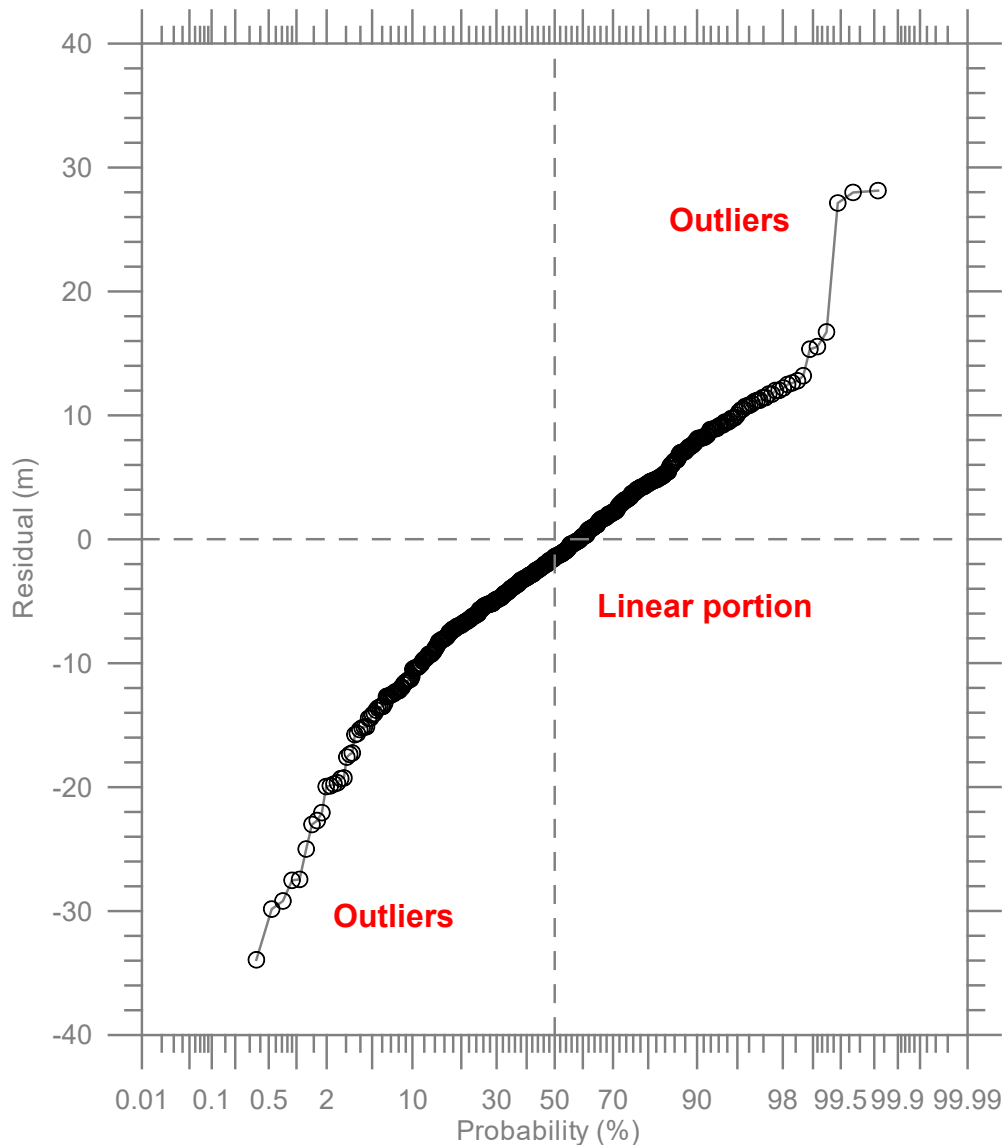


Figure 2. Cumulative probability plot of model residuals

5. Other considerations in assessing whether a model is fit for purpose

- A posting of the residuals on contours of calculated water levels should confirm that there are no areas of systematic mismatch. The patterns of residuals should suggest that there is an inherent variability in the calibration targets that cannot be resolved in the analysis. In the case of fractured-rock, for example, we might expect that in some locations relatively large positive and negative residuals are located close together.
- Inferred hydraulic conductivities should be constrained by geologic interpretations, and consistent with the results of hydraulic testing.
- The cumulative recharge should be consistent with inferred surface water baseflow (minus surface and groundwater consumption). In general, the recharge should be a fraction of the water surplus.

6. References

- Anderson, M.P., and W.W. Woessner, 1992: **Applied Groundwater Modeling**, Academic Press, Inc., San Diego, California.
- ASTM, 1993: Standard Guide for Comparing Ground-Water Flow Model Simulations to Site-Specific Information, ASTM Standard Guide D5490-93, American Society for Testing and Materials, Philadelphia, Pennsylvania.
- Hill, M.C., 1998: Methods and Guidelines for Effective Model Calibration, U.S. Geological Survey, Water-Resources Investigation Report 98-4005.
- Hill, M.C., and C.R. Tiedeman, 2002: Weighting observations in the context of calibrating groundwater models, in *Calibration and Reliability in Groundwater Modeling: A few steps closer to reality*, ModelCARE'2002, Prague, Czech Republic, June 17-20, 2002, International Association of Hydrologic Sciences, Publication No. 277, pp. 196-203.
- Loague, K., and R.E. Green, 1991: Statistical and graphical methods for evaluating solute transport models: Overview and application, *Journal of Contaminant Hydrology*, 7, pp. 51-73.
- Spitz, K., and J. Moreno, 1996: **A Practical Guide to Groundwater and Solute Transport Modeling**, John Wiley & Sons, Inc., New York, New York.
- Woessner, W.W., and M.P. Anderson, 1990: Setting calibration targets and assessing model calibration – Room for improvement: An example from North America, in *Calibration and Reliability in Groundwater Modeling*, K. Kovar (ed.), International Association of Hydrologic Sciences, Publication No. 195, pp. 279-288.
- Woessner, W.W., and M.P. Anderson, 1992: Selecting calibration values and formulating calibration targets for ground-water flow simulations, in *Proceedings of the Fifth International Conference on the Use of Models to Analyze and Find Working Solutions to Ground Water Problems*, National Ground Water Association, Ground Water Management Series Book 9, pp. 199-212.
- Zheng, C., and G.D. Bennett, 2002: **Applied Contaminant Transport Modeling**, Second edition, Wiley-Interscience, New York, New York.

Model acceptance criteria per Spitz and Moreno (1996)

Christopher J. Neville
S.S. Papadopoulos & Associates, Inc.
Last update: May 27, 2008

Spitz and Moreno (1996; p. 240-241) write:

“Common examples of such testing are variance between predicted and observed data should be less than about 10% of the range of observation, standard deviation between predicted and observed data should lie in the range 0.7 to 1.0 ...”

These model acceptance criteria are presented without reference to any examples from the literature. Immediately troubling is that they do not make sense as stated:

- The variance of residuals has units of L^2 (e.g., ft^2 or m^2), while the range of observations has units of L (ft or m). It is not possible to express a ratio of the variance and range in dimensionless form, that is, as a quantity like “10%”; and
- The standard deviation of residuals has units of L . It is not possible to express this quantity in a dimensionless form that ranges from 0.0 to 1.0.

At the MODFLOW 2003 conference in Golden, Colorado, I had the opportunity to meet Joanna Moreno. At that time she indicated that no errata sheets had been published for the Spitz and Moreno textbook. She did admit that the acceptance criteria as stated do not make sense. The closest I have come to an erratum is an e-mail exchange in which Ms. Moreno inserted comments into my e-mail query.

Ms. Moreno never did provide sources for the acceptance criteria that she presented in her book. In my opinion it is appropriate to interpret them as rules-of-thumb, rather than as formal criteria for evaluating a model. In spite of my reservations, I note that the “10% rule” is presented frequently as if it has a firm basis, and is taken as the sole criterion for the reliability of a calibrated model.

TO: CHRISTOPHER NEVILLE
 Fr: J. Moreno



"Christopher Neville"
 <cneville@sspa.com>

10/14/2003 09:41 AM

To: <joanna_moreno@urscorp.com>
 cc:
 Subject: Spitz and Moreno: Assessing model calibration

Dr. Moreno:

I very much enjoyed meeting you last month at the MODFLOW 2003 conference. As I indicated during our conversation, I am a great admirer of your textbook. At the time of the conference I forgot to ask some questions about a section of your book that has been puzzling me for a while. I hope you don't mind me asking those questions now.

On page 240-241, you write:

Common examples of such testing are variance between predicted and observed data should be less than about 10% of the range of observations, standard deviation should lie in the range of 0.7 to 1.0 depending on the number and quality of data points, and bias between predictions and observations should be random rather than systematic.

Handwritten notes:
 - A circle around "variance" with an arrow pointing to "RMSE".
 - A circle around "standard deviation" with an arrow pointing to "correlation coefficient".

The guideline regarding the variance of the residuals is not physically meaningful. In the context of a flow model calibration, the variance of the residuals will have units of length², while the units of the range of observations will be length. What two quantities did you intend to be comparing?

Handwritten: RMSE < 10% (range)

The guideline regarding the standard deviation is not immediately interpretable, as the standard deviation is not dimensionless (it will have units of length as well). Did you mean some sort of normalized standard deviation? If so, what would be the normalizing measure?

And finally, were these suggestions based on work published elsewhere? *Handwritten:* () ← still looking

I thank you in advance for your time in answering these questions.

Christopher J. Neville
 S.S. Papadopoulos & Associates, Inc.
 90 Frobisher Drive, Unit 2B
 Waterloo, Ontario
 N2V 2A1
 t: (519) 579-2100
 f: (519) 579-9779