Journal of Hydrology (2007) 338, 57–62

available at www.sciencedirect.com

ScienceDirect

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

An invalidation test for predictive models

W.E. Bardsley *, J.M. Purdie

University of Waikato, Department of Earth and Ocean Sciences, Private Bag 3105, Hamilton 3240, New Zealand

Received 14 January 2006; received in revised form 15 February 2007; accepted 20 February 2007

KEYWORDS
Model invalidation;
Permutation test;
Hydrological model;
Significance level

Summary The standard means of establishing predictive ability in hydrological models is by finding how well predictions match independent validation data. This matching may not be particularly good in some situations such as seasonal flow forecasting and the question arises as to whether a given model has any predictive capacity. A model-independent significance test of the presence of predictive ability is proposed through random permutations of the predicted values. The null hypothesis of no model predictive ability is accepted if there is a sufficiently high probability that a random reordering of the predicted values will yield a better fit to the validation data. The test can achieve significance even with poor model predictions and its value is for invalidating bad models rather than verifying good models as suitable for application. Some preliminary applications suggest that test outcomes will often be similar at the 0.05 level for standard fit measures using absolute or squared residuals. In addition to hydrological application, the test may also find use as a base quality control measure for predictive models generally.

Introduction

It is usual for hydrological models to be constructed with a view to having predictive ability for subsequent practical applications. A necessary condition in this regard is that the model be shown capable of generating predictions not too far removed from the values of an independent validation data set. The model may then be further evaluated with respect to a range of criteria which depend on the nature of the model and its area of intended application. Discussions on the nature of this validation process can be found, for example, in Anderson and Bates (2001) and Hassan (2004). However, there is no point in persevering with any form of validation when a model has no evident predictive ability (Beven, 2001). That is, a demonstrated lack of predictive ability is a sufficient condition for a hydrological model to be invalidated at the initial phase of investigation.

High levels of predictive ability are usually self-evident but when prediction ability is questionable there is a need for a decision framework by which models may be deemed invalid in the sense of displaying no predictive ability. An obvious approach in this regard is through hypothesis testing whereby the null hypothesis of no predictive ability is accepted or rejected at some significance level on the

* Corresponding author. Tel.: +64 7858 5011; fax: +64 7856 0115.
E-mail address: e.bardsley@waikato.ac.nz (W.E. Bardsley).

0022-1694/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2007.02.019

evidence of how well the model has fared against the validation data. However, despite early recognition for the need for testing models against invalidation (Bredehoeft and Konikow, 1993) there is still a need for general hypothesis testing procedures which would allow invalidation testing of complex predictive models (Refsgaard et al., 2005).

Parametric hypothesis testing procedures relating to model viability include linear regression between observed and predicted values (Flavelle, 1992), assignment of parametric error structures to specific model configurations (Luis and McLaughlin, 1992) and the factor-of-f test to check whether model predictions fall within specified confidence internals (Parrish and Smith, 1990). Non-parametric techniques include the non-parametric equivalent of the factor-of-f test (Zacharias et al., 1996) and permutation tests for the specific case of models derived from variable selection in linear regression (Lindgren et al., 1996). Zacharias et al., 1996 make mention of non-parametric hypothesis testing through generating distributions of goodness of fit indices by bootstrap resampling. Robinson and Froese (2004) reference other tests in various contexts and also make the important point that the usual significance tests have the undesirable feature of model non-invalidation as the default state. They then propose an alternative approach using equivalence tests by which model invalidation becomes the null hypothesis.

We present here a particularly simple and general non-parametric test of invalidation of predictive hydrological models by defining predictive invalidation in a random permutation sense. That is, a model is deemed to be invalid if there is an unacceptably high level of probability that random pairings of predicted and observed values will yield a better goodness of fit measure than was obtained from the original prediction sequence. This is arguably the most basic of all possible tests of model predictive ability but to our knowledge this approach has not been previously advocated in the environmental sciences. The permutation test is easy to carry out and has the advantage of being independent of the nature of the model which generated the predictions. For example, the predicted values might be obtained as weighted averages of individual predictions from a number of different non-linear models. A further useful feature of the test is that the null hypothesis is for model invalidation and so avoids the concerns raised by Robinson and Froese (2004).

Test procedure

It is assumed that an independent validation data set of n recorded values is available which is representative of the kind of conditions where the model might be subsequently applied in a predictive capacity. For each recorded observation in the validation data there is an associated model-predicted value and these observed and predicted pairings collectively yield some numerical fit measure Z, obtained from a goodness of fit expression. The fit expression is utilised here simply as a means of comparison and Z is not an estimate of some unknown true value.

A permutation test is then carried out by way of random pairings of the predicted and observed values. The test can be applied with any goodness of fit expression, which would be chosen so as to best detect the validation data matchings of most interest.

Once an appropriate goodness of fit measure has been selected for application to the validation data, the test procedure is to carry out k random permutations of the n model-predicted values to create k new sets of n observed-predicted pairings, yielding k simulated goodness of fit values. The test quantity p is defined as the proportion of the k simulated fit values which are equal to or better than Z. ''Equal to or better'' here denotes either $\leqslant Z$ or $\geqslant Z$ depending on whether the fit expression is an increasing or decreasing function of fit. If it happens that p is unacceptably large (for example, $p > 0.05$) then the null hypothesis of an invalid predictive model is accepted.

The value of k is determined by the accuracy required for p, which is a binomial proportion. Given that interest is restricted to p values of 0.05 or less, $k = 100{,}000$ will yield p to an accuracy of at least two decimal places. Obtaining the k fit values is a simple procedure and requires only a routine for generating random permutations of the integers $1, 2, \ldots, n$, which serve as array indices of the n reshuffled predicted values. For example, the Matlab function RANDPERM can be used to generate random integer permutations.

If it happens that the number of observed and predicted pairings is small, say $n \leqslant 11$, then an alternative approach to carrying out k random permutations is to explicitly evaluate fit values for all $n!$ possible orderings of the predicted values. The p value so obtained is then exact.

Discussion

The outcome of the test is generally not independent of the selected goodness of fit index. For example, a fit measure using squared deviations is likely to reflect how well the model predictions match the more extreme values in the validation data set, while absolute deviations put more uniform weight over the data range. It is therefore possible that a given validation data set might yield significance for one goodness of fit measure and non-significance for another. However, if one fit expression is simply an increasing or decreasing function of another then both expressions will yield the same test results. A discussion of various fit expressions and their properties can be found in Legates and McCabe (1999) and Coffey et al. (2004). Some example comparisons are given in the next section of test outcomes using fits raising fit residuals to different powers.

One situation where the test is unlikely to detect an invalid model is when a standard goodness of fit measure is incorrectly applied to validation data with systematic spatial or temporal variation. For example, a time series validation set might contain a strong seasonal signal and the model only has to forecast the seasonal means to achieve high significance in the test. The correct validation data set here are the residuals from the seasonal means rather than the recorded data itself. Or, alternatively, a more appropriate index of fit could be selected which can explicitly incorporate seasonal effects (Legates and McCabe, 1999).

The permutation test forces a dichotomy upon a predictive model such that it is deemed to be either invalidated or not invalidated, as opposed to invalidated or validated. It is

entirely possible that highly significant p values will sometimes be associated with validation fits which are too poor to be of any practical application. This is well illustrated by noting that the p value is not changed if an arbitrary constant is added to the validation data. Also, near-zero values of p will be standard in rainfall-runoff models and other situations where the predicted values tend to be strongly correlated. Random permutation in this situation is most unlikely to generate a random sequence with similar apparent data correlation so the original goodness of fit value is unlikely to be exceeded. On the other hand, a non-significant value of p can be taken as model failure regardless any form of association between the original predicted values. We suggest therefore that evaluation of p might be the first step in the validation process of all environmental predictive models, recognising that generating near-zero p values will be a formality for many cases.

The concept of ''model invalidation'' is somewhat terminal but is always with respect to a given validation data set. It could happen that a prediction model yields non-significant p values only for physical situations producing validation date sets with certain characteristics. In this case the test serves as a formal mechanism to limit the range of application of the model. Similarly, it would be premature to deem a model to be invalidated when non-significance arises because circumstances permit only a small value of n. The interpretation in this case is rather that the validation data cannot be given as evidence that the model might be useful. However, this conclusion could be reversed later if a larger validation set gives significantly small p values. The alternative scenario is for p values to remain large as n increases, showing the predictive model has no evident value.

The invalidation test should find particular application in poor prediction situations such as seasonal discharge forecasting where it might be of scientific interest to determine whether model prediction capacity is present at all. It might happen that the test in fact identifies a number of models with significant p values, providing a starting point for further model development with a view to obtaining a final single model with best predictive ability.

Another area of possible application could be for validation subset analysis. For example, a rainfall-runoff model could have a validation data subset comprised of river flood peaks above some threshold magnitude. It may be that this subset yields a non-significant p value despite good fits being obtained to the flow hydrograph as a whole. This would have obvious implications if the main purpose of the model is for flood peak forecasting.

Examples

For illustrative purposes we utilise the coefficient of efficiency as the goodness of fit measure. This can be written in generic form as (Legates and McCabe, 1999)

$$E_c = 1.0 - \frac{\sum_{i=1}^{n}|O_i - P_i|^c}{\sum_{i=1}^{n}|O_i - \overline{O}|^c} \quad (1)$$

where c is a positive integer. Specifically, we apply (1) to a selection of examples for the particular case of E_2, which is the form of the coefficient of efficiency involving squared residuals used in numerous hydrological goodness of fit evaluations. We then tabulate the effect on p when c takes on some values other than 2. All p values listed here are accurate to at least the number of decimal places indicated.

The first example is derived from a study evaluating the predictive ability of a range of different hydroclimatic models for season-ahead river inflow forecasting for Lake Pukaki in New Zealand. One example where the null hypothesis of no predictive ability is accepted is shown in the validation set plotted in Fig. 1. The p value of 0.06 is only just above the 0.05 significance level and it might be anticipated that the model rejection could have altered to model acceptance if it had not been for the one particularly bad inflow forecast in 1976. In fact, carrying out the test with 1976 removed makes very little difference with E_2 reducing slightly to 0.24 and p remaining at 0.06.

If a larger data set is synthesised by simply repeating this validation data to give $n = 20$ then E_2 remains unchanged at

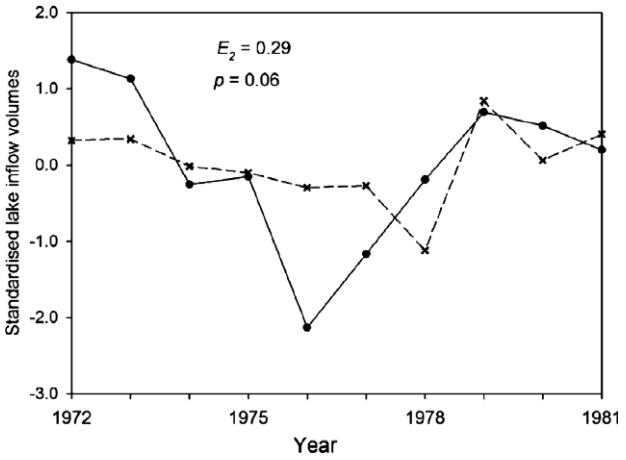

Figure 1 Recorded (solid line) and predicted (dashed line) spring season inflow volumes to Lake Pukaki, New Zealand (standardised volume units) for a 1972—1981 validation period. Predicted inflows are obtained from a season-ahead hydroclimatic forecasting model.

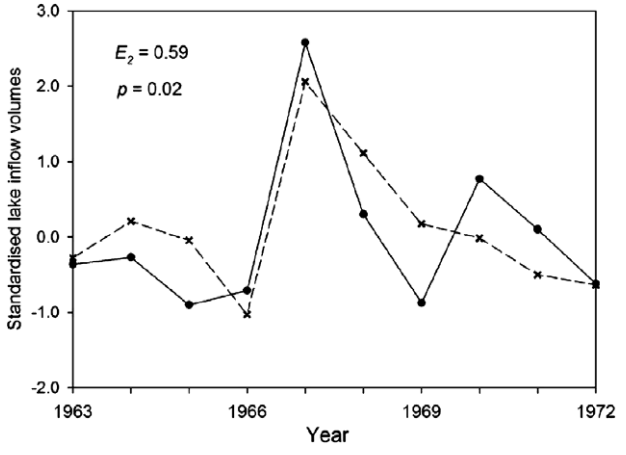

Figure 2 Recorded (solid line) and predicted winter inflow volumes to Lake Pukaki, New Zealand (standardised volume units) for a 1963—1972 validation period. Predicted inflows are obtained from a season-ahead hydroclimatic forecasting model.

0.29 but p reduces to a significantly small value of 0.01. The enlarged data set is only a contrivance of course and contains no further information, but serves to illustrate the increased power of the test to detect predictive ability as n increases.

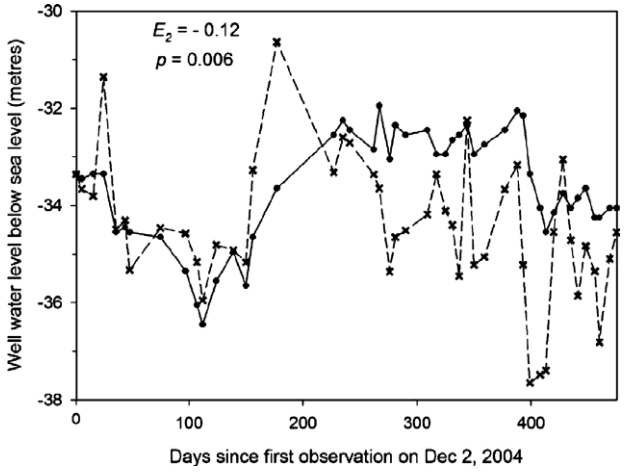

Figure 3 Time series of recorded (solid line) and predicted water levels during pumping from a coastal well near Whangamata, New Zealand.

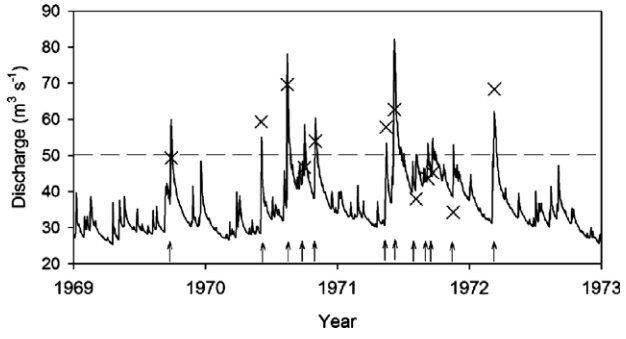

Figure 4 Four-year daily discharge validation data set for peak flood discharges on the Tarawera River, New Zealand. Arrows indicate 12 flood peaks exceeding 50 $m^3 s^{-1}$ and crosses are rainfall-runoff model predictions of the discharge peaks.

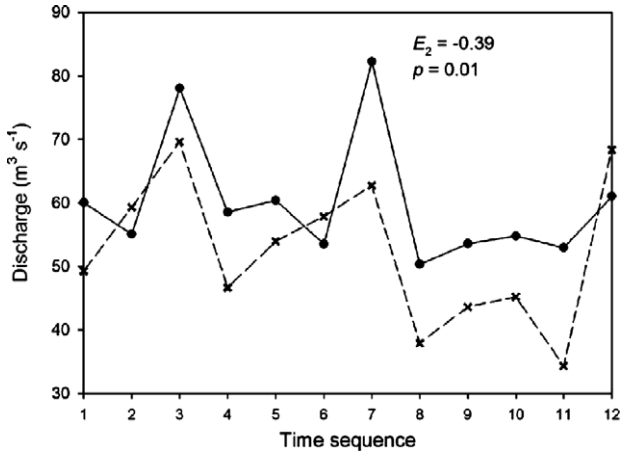

Figure 5 Recorded (solid line) and predicted discharges for the 12 flood peaks of Fig. 4, plotted in time order.

Fig. 2 shows a different seasonal inflow forecasting model applied to a different validation period where the p value of 0.02 indicates the null hypothesis of no predictive ability is rejected at the 5% level.

Fig. 3 illustrates a validation example for well water level predictions derived from an empirical regression-based model based on well pumping rates. It is evident that the predictive model is not particularly helpful in this case because the negative value of E_2 means that the observation mean value gives a better prediction than the model. The p value of 0.006 is nonetheless highly significant as a consequence of similarities between the patterns of the observed and predicted time series. This would give encouragement to persevere with the regression model and hopefully introduce further terms which improve the goodness of fit.

Fig. 4 shows four years of daily flow record from the Tarawera River, New Zealand. This serves as the validation data of a class exercise evaluating the application of a specific rainfall-runoff model (Bardsley and Liu, 2003). The daily predicted values are not plotted for the sake of clarity, but the high degree of temporal correlation resulted in none of 100,000 random permutations exceeding the original calculated E_2 fit value of 0.55. This situation will be common to many rainfall-runoff model validations and raises the question of how p should be reported, since specifying $p = 0$ is incorrect. We suggest instead presenting the upper 95% upper confidence bound to p whenever k random permutations yield no goodness of fit values better than the original. That is, an upper bound to p is given as (Louis, 1981)

$$p < 1 - 0.05^{1/k} \tag{2}$$

For example, 100,000 randomisations with zero better fits gives $p < 0.3 \times 10^{-4}$.

The test is likely to provide more interesting results for rainfall-runoff models when checking the ability of a model to predict uncorrelated magnitudes of high or low flow extremes. For example, Fig. 4 also plots the predicted values of the recorded 12 flood peaks which exceed 50 $m^3 s^{-1}$ over the four years (plotted separately in Fig. 5 as a time se-

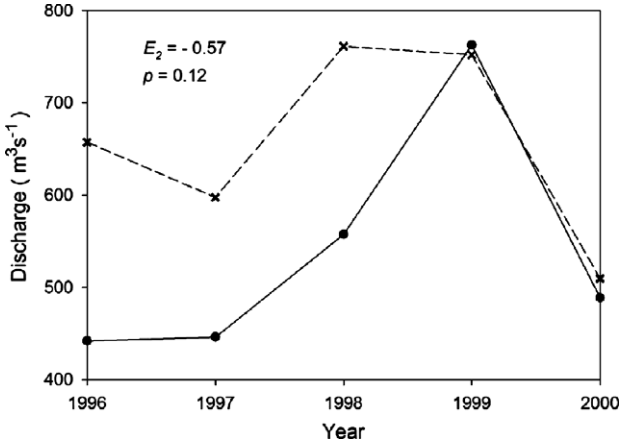

Figure 6 Recorded (solid) and predicted mean annual discharges for Tangnaihai Station on the upper Yellow River (China). Predicted values are obtained from a composite GCLM and SWAT model.

Table 1 Selected E_c fit values applied to the example data sets, with corresponding p values in brackets

	[1]	[2]	[3]	[4]	[5]	[6]
$E_{0.5}$	0.08 (0.09)	0.08 (0.09)	0.13 (0.03)	−0.13 (0.003)	−0.38 (0.14)	−0.07 (0.23)
E_1	0.16 (0.07)	0.14 (0.07)	0.26 (0.02)	−0.39 (0.003)	−0.49 (0.03)	−0.25 (0.20)
E_2	0.29 (0.06)	0.24 (0.06)	0.59 (0.02)	−1.2 (0.006)	−0.39 (0.01)	−0.57 (0.12)
E_3	0.40 (0.07)	0.38 (0.05)	0.83 (0.02)	−2.6 (0.013)	−0.17 (0.004)	−0.68 (0.10)

[1] From data of Fig. 1; [2] from data of Fig. 1 excluding the 1976 year; [3] from data of Fig. 2; [4] from data of Fig. 3; [5] from data of Fig. 5; [6] from data of Fig. 6.

quence). For this subset of validation data the model is evidently not particularly helpful because it tends to under-predict the flood peaks and gives a negative value of E_2. However, the low p value of 0.01 suggests the model has captured some aspects of the flood generation process and could perhaps be further developed.

The final example illustrates the use of the test as a check on the extent to which a small validation data set can be presented as evidence in support of a predictive model. In this context, Fig. 6 shows the validation data (n = 4) of a global climatic model coupled with a SWAT model to give predictions of mean annual flow on the upper Yellow River (Xu and Zhao, 2006). Applying the test, the resulting p value of 0.12 is in fact not sufficiently small to confirm predictive ability but might be considered small enough to encourage putting together a larger validation data set as a more rigorous check of the model. There is also need to improve the model's low fit value.

All the examples considered here have been based on the single goodness of fit statistic E_2 and the question arises as to how p might vary for different fit measures. As mentioned earlier, different fit measures emphasise different data fit aspects so some validation sets will result in greater differences among the p values than others. A full investigation is left for further work but a preliminary indication of p variation is given in Table 1, which shows the p values in the examples of this paper for the fit measures $E_{0.5}$, E_1, and E_3. It is not implied that indices like $E_{0.5}$ or E_3 should ever be used as practical fit measures but it is of interest to see how p may change over this range. It is encouraging to note that at the 0.05 significance level the test conclusions would remain unchanged for both E_1 and E_2 and the associated p values tend to be similar. However, this evident robustness of the p values needs confirmation by subsequent application to a range of validation data sets. The greater difference in the square root and third power of the absolute residuals in $E_{0.5}$ and E_3 tends to give rise to some larger differences between their respective p values.

Conclusion

Random permutation of model-generated predicted values provides a simple and general means for testing the null hypothesis of a model's inability to predict a validation data set. The permutation test is a simple acceptance/rejection procedure and further investigation would be required for possible explanations as to why the test gave a particular outcome. The test is likely to find most application in exploratory analysis seeking to identify models which hold some promise of predictive ability in situations were accurate prediction is inherently difficult. The test could also be used as part of a standard statement of quality control for predictive models in general, with strong rejection of the null hypothesis expected to be the norm for models with reasonable predictive capability over large validation data sets.

Acknowledgements

We thank Meridian Energy Ltd. for providing the Lake Pukaki inflow data. The well water level data were provided by the Thames Coromandel District Council and the groundwater predictive model was developed by Neil Jelley. The Tarawera River flow data were provided by Environment Bay of Plenty. We are particularly grateful to Dr. Zongxue Xu of Beijing Normal University for making available his Yellow River discharge validation data set. Finally, we are appreciative of helpful comments by anonymous referees in the review process.

References

Anderson, M.G., Bates, P.D., 2001. Model Validation: Perspectives in Hydrological Science. Wiley, Chichester.

Bardsley, W.E., Liu, S., 2003. An Approach to Creating Lumped-parameter Rainfall-runoff Models for Drainage Basins Experiencing Environmental Change, 281. IAHS Publ., pp. 67–74.

Beven, K., 2001. On explanatory depth and predictive power. Hydrol. Process. 15, 3069–3072.

Bredehoeft, J.D., Konikow, L.F., 1993. Ground-water models: validate or invalidate. Ground Water 31, 178–179.

Coffey, A.E., Workman, S.R., Taraba, J.L., Fogle, A.W., 2004. Statistical procedures for evaluating daily and monthly hydrologic model predictions. Trans. ASAE 47, 59–68.

Flavelle, P., 1992. A quantitative measure of model validation and its potential use for regulatory purposes. Adv. Water Resour. 15, 5–13.

Hassan, A.E., 2004. Validation of numerical ground water models used to guide decision making. Ground Water 42, 277–290.

Legates, D.R., McCabe, G.J., 1999. Evaluating the use of Goodness of Fit measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35, 233–241.

Lindgren, F., Hansen, B., Karcher, W., Sjöström, M., Eriksson, L., 1996. Model validation by permutation tests: applications to variable selection. J. Chemom. 10, 521–532.

Louis, T.A., 1981. Confidence intervals for a binomial parameter after observing no successes. The American Statistician 35, 154.

Luis, S.J., McLaughlin, D., 1992. A stochastic approach to model validation. Adv. Water Res. 15, 15—32.

Parrish, R.S., Smith, C.N., 1990. A method of testing whether model predictions fall within a prescribed factor of true values, with an application to pesticide leaching. Ecol. Modell. 51, 59—72.

Refsgaard, J.C., Henriksen, H.J., Harrar, W.G., Scholten, H., Kassahun, A., 2005. Quality assurance in model based water management — review of existing practice and outline of new approaches. Environ. Modell. Softw. 20, 1201—1215.

Robinson, A.P., Froese, R.E., 2004. Model validation using equivalence tests. Ecol. Modell. 176, 349—358.

Xu, Z., Zhao, F., 2006. Comparison between statistical downscaling and delta methods for the simulation of streamflow in headwater of the Yellow River basin. Paper presented at the American Geophysical Union Western Pacific Meeting, Beijing, Abstract H21A-04.

Zacharias, S., Heatwole, C.D., Coakley, C.W., 1996. Robust quantitative techniques for validating pesticide transport models. Trans. ASAE 39, 47—54.