A goodness of fit measure related to r^2 for model performance assessment

W E Bardsley, School of Science
University of Waikato
New Zealand
e.bardsley@waikato.ac.nz

ABSTRACT
Checking the predictive worth of an environmental model inevitably includes a goodness of fit metric to quantify the degree of matching to recorded data, thereby giving a measure of model performance. Considerable analysis and discussion has taken place over fit indices in hydrology but a neglected aspect is the degree of communicability to other disciplines. It is suggested that a fit index is best communicated to colleagues via reference to models giving unbiased predictions, because unbiased environmental models are a desirable goal across disciplines. That is, broad recognition of a fit index is aided if it simplifies in the unbiased case to a familiar and logical expression. This does not hold for the Nash-Sutcliffe Efficiency E which reduces to the somewhat awkward unbiased expression $E = 2 - 1/ r^2$, where r^2 is the coefficient of determination. A new goodness of fit index V is proposed for model validation as $V = r^2/(2-E)$, which simplifies to the easily-communicated $V = r^4$ in the unbiased case. The index is defined over the range $0 \le V \le 1$ and it happens that $V < E$ for larger values of E. Some synthetic and recorded data sets are used to illustrate characteristics of V in comparison to E.

INTRODUCTION

Goodness of fit indices are useful metrics whereby a single number is used to summarise how well a model performs in matching a set of validation data, recognising at the same time that proper model evaluation should always include multiple criteria which will vary with the nature of the model (Biondi et al, 2012). Many different fit measures and validation concepts have been developed and discussed over the years in hydrology and other subject areas – see, for example, Legates and McCabe (1999), Biondi et al (2012), Pushpalatha et al (2012), Krause et al (2005), Dawson et al (2007), Coffey et al (2004), Lin et al (2002), Tedeschi (2006), Criss and Winston (2008). A useful overview on fitting measures and methodology is given by Bennett et al (2013).

The hydrology community have tended to favour as a validation fit measure the Nash-Sutcliffe Efficiency E:

$$E = 1 - \frac{\sum (O_i - P_i)^2}{\sum (O_i - \bar{O})^2} \qquad -\infty < E \le 1 \tag{1}$$

as proposed by Nash and Sutcliffe (1970). The terms O_i and P_i here denote respectively the observed data and model-predicted values. E is a comparison against the mean value as a baseline predictor. Many other baselines could be used as well, depending on the situation (Schaefli and Gupta, 2007; Seibert, 2001).

There has been much analysis and discussion of E since its introduction and relevant work includes McCuen et al (2006), Gupta et al (2008), Criss and Winston (2008), Jain and Sudheer (2008), Gupta and Kling (2011) and Ritter and Muñoz-Carpena (2013). An important point noted by Schaefli and Gupta (2007) is that E is unfamiliar in the wider field of environmental sciences. Hydrology has many connections to other disciplines and it would seem that this issue of communication could be given more attention.

Fit indices differ in their mathematical expressions, depending on the means of quantification of misfit due to bias effects (systematic departure from the 1:1 line) and to random scatter. The environmental sciences aspire to develop models which are at least approximately unbiased and therefore the degree of communicability of a fit index might be measured by its degree of familiarity when simplified for the unbiased case. The purpose of this brief communication is to present a new goodness of fit index V which has the attribute of being immediately communicable in its unbiased form as r^4, which is simply the square of the coefficient of determination.

V is proposed here in the sense of a performance assessment of a single defined model as applied to a given validation data set. It is not intended as a tool for model identification or for making a statement as to general applicability of any specific model. Nor is it proposed that *V* should necessarily be used as an index in calibration because *V* in calibration optimisation will have the same issues of component aliasing as previously identified for *E* (Gupta et al, 2009). Also, it may happen that *V* in calibration has a somewhat more constrained range than for validation, as is the case for *E* (Gupta and Kling, 2011).

FIT INDEX

The new fit index can be defined conveniently in terms of *E* and r^2 as:

$$V = r^2 / (2 - E) \qquad 0 \le V \le 1 \qquad (2)$$

where r^2 is the coefficient of determination with respect to the linear regression relation between the observed data and the model-predicted values. The index could be thought of as the product of r^2 and the term 1/(2-*E*) which rescales *E* to the 0,1 interval.

An extreme special case where *V* would fail is if r^2 = 1 exactly and the regression line between observed and model-predicted values is close to the x-axis. The model concerned should be rejected of course because all model predictions are 0.0 for practical purposes. In this situation $E \approx 0$ but *V* has the unreasonably high value of 0.5. In practical applications, however, there will always be some scatter about the regression line which, if near the x-axis, would give $r^2 \approx 0$ and therefore $V \approx 0$.

Table 1 lists the unbiased-case expressions for some dimensionless fit indices. It is evident that *E* simplifies to a somewhat awkward function of r^2 which is not so amenable to ease of communication. Similar comments apply for the equivalent unbiased expression

Symbol	Expression (unbiased case)	Fit Measure
E	$2 - 1/r^2$	Nash-Sutcliffe Efficiency
C_{2M}	$2r^2 - 1$	*E* defined over the -1 +1 interval
r_C	r	Concordance correlation coefficient
ωr^2	r^2	Weighted coefficient of determination
V	r^4	Proposed new fit measure

Table 1. Unbiased-case expressions for selected dimensionless fit indices.

of C_{2M} , which is *E* mapped to the -1, +1 interval (Mathevet et al 2006). Lin (1989) introduced the concordance correlation coefficient r_C which has been widely adopted as a fit index in medical statistics and has found some application in hydrology (Meek et al 2009). The unbiased expression for r_C is just the correlation coefficient *r* which is immediately recognisable. Despite this familiarity advantage however, r_C has not gained popularity in hydrology. This may be because the complete r_C expression is somewhat complicated. In contrast, the weighted coefficient of determination ωr^2 (Krause et al, 2005) is certainly simple and is just r^2 in the unbiased case. However, ωr^2 is of limited value as a practical index because the weight ω represents only the degree of prediction mismatch arising from proportionality differences, neglecting over- or under-prediction bias.

In addition to ease of communication, *V* has some advantage over *E* in that *V* gives smaller values that *E* for situations of good fit, keeping in mind that *E* is criticised from time to time for yielding somewhat larger values than seems desirable. Specifically, for any $r > 0$ there is equality of *E* and *V* for a particular value *E** defined by:

$$E^* = 1 - r\sqrt{r^{-2} - 1} \qquad (3)$$

If $E > E^*$ then $V < E$ and vice versa (Figure 1).

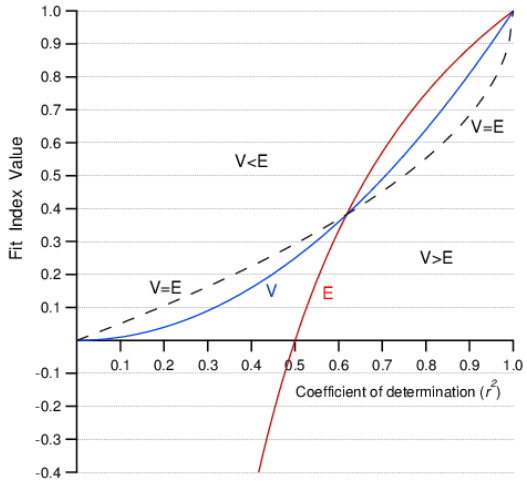

Figure 1. Comparison of V and E as a function of r^2. Dashed line denotes $V = E$ and is a plot of E^* in Equation 3. The region above the dashed line represents $V < E$. Blue and red lines are respectively $V = r^4$ and $E = 2 – 1/ r^2$, which are the unbiased cases for V and E.

Beyond noting the ease of communication aspect, it is not the intention of this brief paper to put forward an objective case for general use of V by way of detailed mathematical or data-based comparison with other fit indices. Selection of an index will in the end be a matter of personal choice. However, some sense of the behaviour of V in comparison to E can be seen in the following section with respect to some selected data sets.

APPLICATION TO SYNTHETIC AND RECORDED DATA SETS

Figures 2-7 illustrate V and E applied to synthetic data to show some specific fit situations. Figures 2-4 show V and E for unbiased models with progressively decreasing degrees of fit. That is, Figures 2-4 all have a = 0, b=1, and $V = r^4$, where a and b are respectively the intercept and gradient of the linear regression line between observed and predicted values. The indices in the plots are given to three decimal places to enable comparisons but two decimal places would be standard. A randomisation-based significance measure p is applied to V but other approaches such as bootstrapping could be used as well (Ritter and Muñoz-Carpena, 2013).

Figures 5-7 illustrate bias effects on V and E. Figure 5 shows an arbitrary scatter of points with no evident association between observed and predicted values. The model here is biased (b = 0.14, a = 12.10) but not statistically significant and the model would be rejected. Figures 6 and 7 give an indication of the response of V to systematic departure from the 1:1 line. Figure 6 illustrates bias in proportionality only (a = 0, b = 1.3). The value of r^2 = 0.92 is clearly too high here as a fit measure (r^2 being a measure of precision but not accuracy) but it is a matter of personal judgement whether V or E best reflects the bias in the predicted values. Figure 7 illustrates the effect of bias arising from a displacement effect only (a = 10, b =1). Again, it is personal preference whether V or E better represents the considerable degree of bias in this example.

Figure 8 shows a recorded hydrograph segment where the “model” is simply an exponential curve fitted to a more extended portion of the discharge data. The corresponding scatter plot is shown in Figure 9. Figures 9 and 10 both show the mean as the better predictor of the data, but with V values indicating the model in Figure 10 giving better fit to data than the model of Figure 9 achieves for its data. This is despite evident under-prediction bias in the Figure 10 model. Both values of V are statistically significant but the p =0.02 value of Figure 9 may not be reliable because of some degree of serial correlation in the data. Figure 11 shows a better model fit situation for a seasonal forecasting validation set.

The validation sets of Figures 9-11 involve small amounts of data so any model performance conclusion is tentative and there is possibility of change in both V and E with the addition of further data. The small data sets do not well determine the least-squares

best fit line between the observed and predicted data but the discrepancy between V and r^4 is suggestive of model bias in all three cases. It would be useful in fact to present both V and r^2 on scatter plots of observed and predicted values.

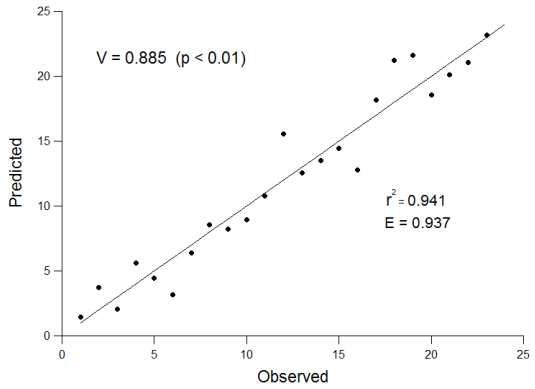

Figure 2. Simulated good fit to data and fit indices (unbiased model). Solid line here and in subsequent figures denotes the 1:1 line. The significance level of V is from randomisation (Bardsley and Purdie, 2007).

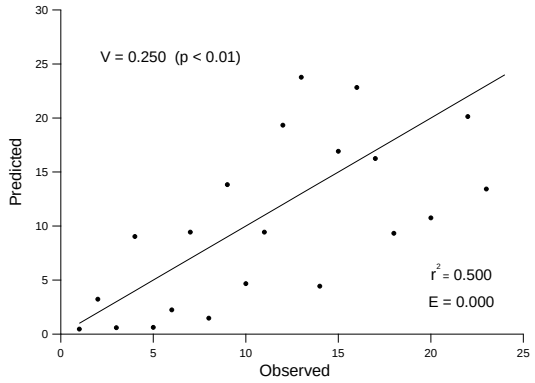

Figure 3. Simulated weak fit to data and fit indices (unbiased model).

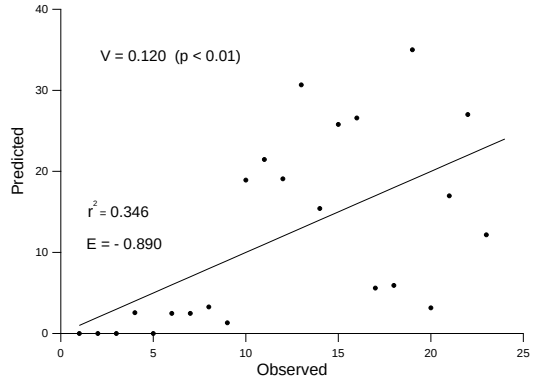

Figure 4. Simulated very poor fit to data and fit indices (unbiased model).

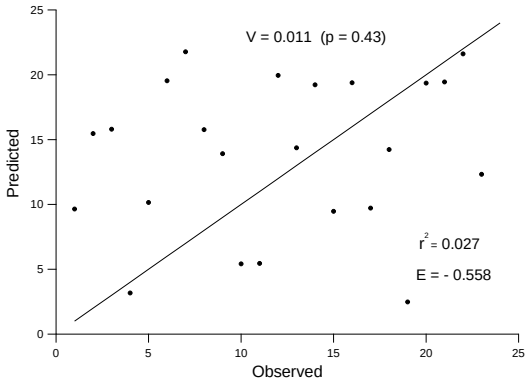

Figure 5. Simulated model failure and fit indices (a = 12.10, b = 0.14).

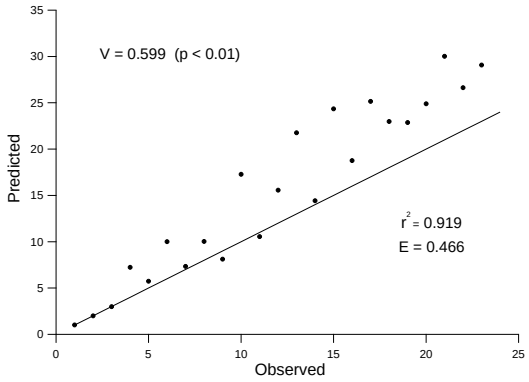

Figure 6. Simulated model fit and fit indices with proportionality bias (a = 0, b = 1.3).

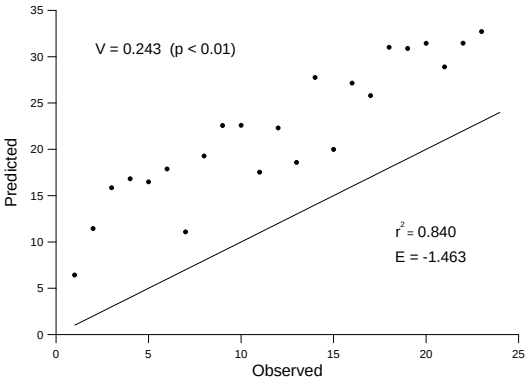

Figure 7. Simulated model fit and fit indices with displacement bias (a = 10, b = 1).

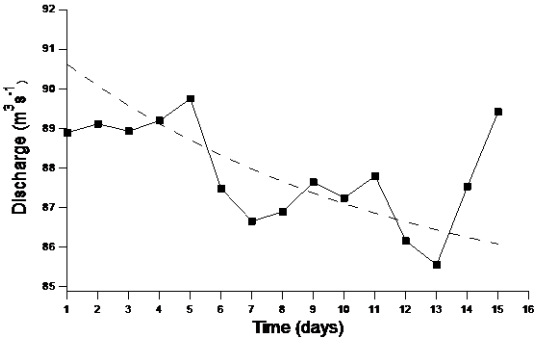

Figure 8. Daily discharge values (from 22/9/1992) of the Kawerau River at Lake Wakatipu outlet, New Zealand. Dashed line is predicted flow from an exponential curve fitted to a longer time series.

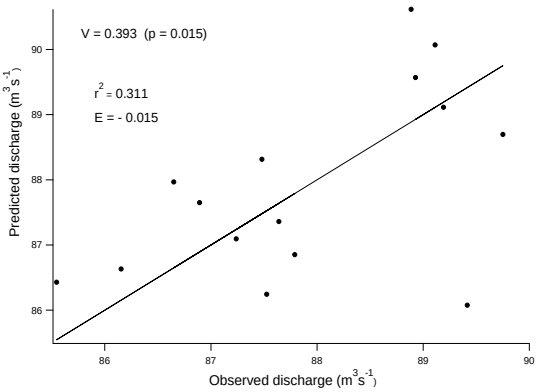

Figure 9. Scatter plot of observed and predicted values from Figure 8.

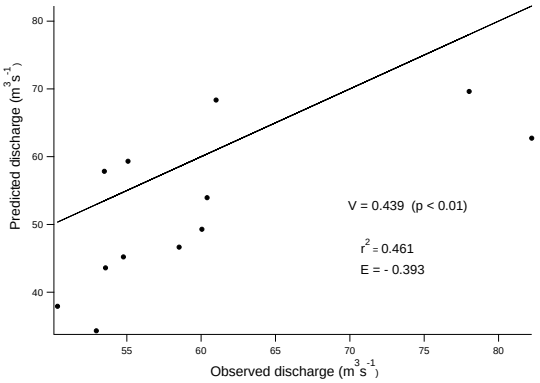

Figure 10. Predictions from a rainfall-runoff model of daily flood magnitudes, Tarawera River, New Zealand. The validation set is recorded flood peaks exceeding 50 m^3s^{-1}. From Bardsley and Purdie (2007).

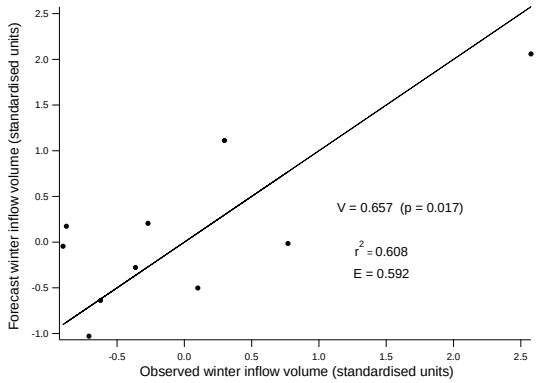

Figure 11. Validation set (1963-97) of a hydroclimatic model forecasting winter inflows to Lake Pukaki, New Zealand. From Bardsley and Purdie (2007).

DISCUSSION AND CONCLUSION

It is well known that fit indices like *V* can yield values which are incorrectly high if applied to models which have a component of discovering the obvious (Schaefli and Gupta, 2007). For example, a model forecasting seasonal rainfall on a per-region basis for wet and dry regions would clearly have no predictive ability if the forecasts were always just the respective regional means, giving *V* = 0 for each region. However, if the regional data were presented as a single data set then the resulting *V* would be an erroneously high value unless the data were first standardised by subtracting regional means from the recorded values.

For example, Legates and McCabe (1999) calculated inflated fit values when comparing various evaporation models because monthly evaporation averages were not first subtracted from the data. The present author apologetically notes a similar error in the context of presenting the fit of a rainfall-runoff model with seasonal river discharge (Bardsley and Liu, 2003).

All fit indices have their advantages and disadvantages in representing validation fit, or lack of fit. One issue relevant to *V* concerns the use of squared deviations, with the implication that the squaring process gives undue weight to the more extreme

observations (Criss and Winston, 2008; Legates and McCabe, 1999). In fact, an argument can be made that such weighting is actually desirable. The reasoning here is that during the calibration process a model gains least experience from the extremes, which are always the most infrequent values. If the calibrated model is able to capture the environmental processes sufficiently well so that extremes are well matched in a validation data set, then it would seem appropriate for a validation fit index to give extra weight to such matching.

V is a convenient fit measure for application to validation data sets. However, no claim is made for it being in some sense better than other indices in terms of summarising model fits by presenting model precision and bias in a single number. The main advantage of *V* is with respect to communicability and in conference presentations, for example, we are not likely to present model prediction plots that are strongly biased. *V* can then be communicated easily as being almost the same as r^4 and therefore a more conservative index than r^2, while at the same time still being a true fit measure and not simply a reflection of model precision. In this spirit of communication, *V* is suggested as a fit index for consideration for use in hydrology and the environmental sciences generally.

REFERENCES

Bardsley WE, Liu S. 2003. An approach to creating lumped-parameter rainfall–runoff models for drainage basins experiencing environmental change. *IAHS Publication* 282: 67–74.

Bardsley WE, Purdie JM. 2007. An invalidation test for predictive models. *Journal of Hydrology* 338: 57–62.

Bennett ND, Croke BFW, Guariso G, Guillaume JHA, Hamilton SH, Jakeman AJ, Marsili-Libelli S, Newham LTH, Norton JP, Perrin C, Pierce, SA, Robson B, Seppelt R, Voinov AA, Fath BD, Andreassian V. 2013. Characterising performance of environmental models. *Environmental Modelling & Software* 40: 1–20.

Biondi D, Freni G, Lacobellis V, Mascaro G, Montanari A. 2012. Validation of hydrological models: Conceptual basis, methodological approaches and a proposal for a code of practice. *Physics and Chemistry of the Earth* 42–44: 70–76.

Coffey AE, Workman SR, Taraba JL, Fogle AW. 2004. Statistical procedures for evaluating daily and monthly hydrologic model predictions. *Transactions of the ASAE* 47: 59–68.

Criss RE, Winston WE. 2008. Do Nash values have value? Discussion and alternate proposals. *Hydrological Process* 22: 2723–2725.

Dawson CW, Abrahart RJ, See LM. 2007. HydroTest: A web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. *Environmental Modelling & Software* 22: 1034–1052.

Gupta HV, Kling H. 2011. On typical range, sensitivity, and normalization of mean squared error and Nash-Sutcliffe efficiency type metrics. *Water Resources Research* 47: W10601, doi:10.1029/2011WR010962.

Gupta HV, Kling H, Yilmaz KK, Martinez GF. 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology* 377: 80–91.

Gupta HV, Wagener T, Liu Y. 2008. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. *Hydrological Processes* 22: 3802–3813.

Jain SK, Sudheer KP. 2008. Fitting of hydrologic models: a close look at the Nash–Sutcliffe Index. *Journal of Hydrologic Engineering* 13: 981–986.

Krause P, Boyle DP, Bäsel F. 2005. Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences* 5: 89–97.

Legates DR, McCabe GJ. 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resources Research 35: 233–241.

Lin LI. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*
45: 255–268.

Lin LI, Hedayat AS, Sinha B, Yang M. 2002. Statistical methods in assessing agreement. *Journal of the American Statistical Association*, 97: 257–270.

Mathevet T, Michel C, Andréassian V, Perrin C. 2006. A bounded version of the Nash–Sutcliffe criterion for better model assessment on large sets of basins. *IAHS Publication* 307: 211–219.

McCuen RH, Knight Z, Cutter, AG. 2006. Evaluation of the Nash–Sutcliffe efficiency index. *Journal of Hydrologic Engineering* 11: 597–602.

Meek DW, Howell TA, Phene CJ. 2009. Concordance correlation for model performance assessment: an example with reference evapotranspiration observations. *Agronomy Journal* 101: 102–108.

Nash JE, Sutcliffe JV. 1970. River flow forecasting through conceptual models. Part I: a discussion of principles. *Journal of Hydrology* 10: 282–290.

Pushpalatha R, Perrin C, Le Moine N, Andréassian V. 2012. A review of efficiency criteria suitable for evaluating low-flow simulations. *Journal of Hydrology* 420–421: 171–182.

Ritter A, Muñoz-Carpena R. 2013. Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. *Journal of Hydrology* 480: 33–45.

Schaefli B, Gupta HV. 2007. Do Nash values have value? *Hydrological Processes* 21: 2075–2080.

Seibert J. 2001. On the need for benchmarks in hydrological modelling. *Hydrological Processes* 15: 1063–1064.

Tedeschi LO. 2006. Assessment of the adequacy of mathematical models. *Agricultural Systems* 89: 225–247.