Groundwater

Technical Commentary/

How Good Is Your Model Fit? Weighted Goodness-of-Fit Metrics for Irregular Time Series

by Raoul A. Collenteur

Introduction

Heads are regularly measured in groundwater monitoring wells all over the world, so also in the hypothetical municipality of Watertown. On July 16, 2011 the hydrogeologist of Watertown starts her monthly visit to the groundwater monitoring wells surrounding the local drinking water pumping station. She measures the head in each monitoring well, as she did every month for the past 10 years. Today, however, marks the last of these regular field visits, as she also installs automatic data loggers to record the head with a daily observation frequency. Fast forward to the year 2020. The municipality of Watertown has had complaints from citizens about declining heads, and wants to know what part of this decline is caused by groundwater pumping.

The hydrogeologist recently read a paper in *Groundwater* on time series analysis (Bakker and Schaars 2019), and decides to try the method for an initial assessment of the drawdown due to pumping. She sets up time series models for each monitoring well and calibrates the models on the entire observation period (2000–2018). To evaluate how well the models perform, she computes the coefficient of determination (R^2) for each model and finds satisfying values for most of them. Looking closer at the results of one of the models, however (see Figure 1), she senses that something is not quite right. To her surprise the visual interpretation shows that the model has a good fit with the observations in the last 8 years of the time series ($R^2 = 0.91$), and a poor fit in the first 10 years ($R^2 = -0.83$). The poor fit for most of the time series, however, was not something she expected from the coefficient of determination that was calculated for the model over the entire calibration period ($R^2 = 0.83$).

The hypothetical story above and the real data shown in Figure 1 are probably familiar to many hydrogeologists. After years of measuring heads by hand at infrequent time intervals (e.g., every month or week), data loggers were installed that started logging the head at higher measurement frequencies (e.g., every day or hour). As a result, many historic head time series are characterized by a mixture of observation frequencies, irregular time steps between observations, and larger data gaps. These characteristics pose challenges to the way we calibrate our models (see, e.g., Bierkens et al. 1999; Yi and Lee 2004; von Asmuth and Bierkens 2005), and to the way we evaluate the model fit. The latter is the topic of this Commentary. For the example data shown in Figure 1, the coefficient of determination computed for the entire period is clearly biased toward the period with more observations. This goodness-of-fit metric is therefore not very informative if one is interested in the average model fit over the entire observation period.

To the best knowledge of the author, taking irregular time steps into account when computing the fit between two time series is not common practice. The aim of this Commentary is therefore to raise awareness of the potential bias in goodness-of-fit metrics when evaluating the fit between two time series with irregular time steps between observations. The use of weighted goodness-of-fit metrics is proposed here as a possible solution. Using the weighted mean and variance, weighted versions of common goodness-of-fit metrics are derived and benchmarked using synthetic head time series. An example application shows how the use of weighted metrics may impact the conclusions we draw about the

Institute of Earth Sciences, NAWI Graz Geocenter, University of Graz, Graz, 8010, Austria; +43 (0)316 380 7406, raoul.collenteur@uni-graz.at

Article impact statement: Weighted goodness-of-fit metrics may help to the evaluate model fit for head time series with irregular time steps.

Received October 2020, accepted May 2021.

doi: 10.1111/gwat.13111

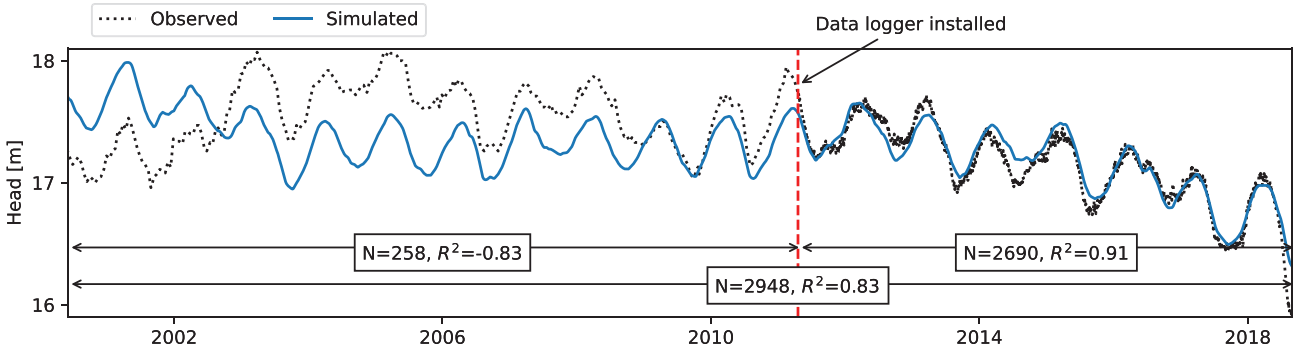

Figure 1. The time series model the hydrogeologist took a closer look at. N is the number of observations in each period.

model fit. This potentially affects how a model is used in subsequent analysis steps. While a time series model was used to illustrate the problem, it is emphasized here that the methods presented in this manuscript are independent of the type of model and the calibration method used to model the heads. Most of the methods presented here are not new, but their application to irregular time series may not be as wide-spread as they perhaps should be.

Weighted Metrics

A large number of goodness-of-fit metrics is available to evaluate the fit between a simulated time series and an observed time series. For an introduction into (hydrological) fit metrics and how to interpret these the reader is referred to Jackson et al. (2019), who provide a thorough review on the topic. Many metrics directly or indirectly depend on the mean (μ) and the sample variance (σ^2) of a time series. For a time series with irregular time steps the standard formulas for these quantities are not applicable, and need to be adapted. This can for example be done by weighting the individual values in the time series. The weighted mean and sample variance for a time series $x(t)$ with irregular time steps (denoted by $\overline{\mu}_x$ and $\overline{\sigma}_x^2$) may be computed as follows:

$$\overline{\mu}_x = \sum_{i=1}^{N} w_i' x_i \tag{1}$$

$$\overline{\sigma}_x^2 = \frac{N}{N-1} \sum_{i=1}^{N} w_i' (x_i - \overline{\mu}_x)^2 \tag{2}$$

where N is the number of observations in the time series $x(t)$, and w_i' [−] are the normalized weights. Note that the sum of the weights should equal unity in this formulation of the mean and the variance and that the weights should be dimensionless. The difficulty now lies in the determination of the weights that can be applied in Equations 1 and 2.

As a pragmatic solution to this problem, we can use the time step between two consecutive observations ($\Delta t_i = t_i - t_{i-1}$) in the computation of the weights. This way a value in a period with few observations receives a relatively high weight, while a value in a period with many observations receives a low weight. To prevent that the first observation after a large data gap receives an unreasonably large weight, the initial weights w_i are computed as follows:

$$w_i = \begin{cases} \Delta t_i & \text{if} \Delta t_i \leq \Delta t_{\max}, \\ \Delta t_{max} & \text{else} \end{cases} \tag{3}$$

where $\Delta t_{\max}$ [T] is a parameter that determines the maximum time step Δt_i that is allowed to be used as a weight. The modeler needs to choose an appropriate value for $\Delta t_{\max}$. For example, if the smallest known observation frequency is once per month, a value of $\Delta t_{\max} = 31$ days may be chosen for this parameter. To obtain normalized weights (w_i') that are dimensionless and sum up to unity, the initial weights (w_i) are divided by their sum as follows:

$$w_i' = \frac{w_i}{\sum_{i=1}^{N} w_i} \tag{4}$$

It is now relatively easy to obtain weighted goodness-of-fit metrics to evaluate the fit between an irregularly observed time series $x(t)$ (e.g., observed heads) and a predicted time series $y(t)$ (e.g., simulated heads). Table 1 shows the equations for six commonly used metrics, adapted to time series with irregular time steps. All of the formulas presented in Table 1 may be derived by substituting the weighted variance and mean into their unweighted versions, or by applying the normalized weights (w_i') to the individuals observations. This also makes it relatively easy to adapt existing software implementations. Note for example that the formula to compute the mean absolute error (MAE) is similar to the formula for the weighted mean, and that the formula for the root mean squared error (RMSE) is closely related to that of the weighted variance. Table 1 includes absolute error metrics (MAE and the RMSE), percentage error metrics (r, R^2, and the explained variance percentage) and a forecast error metric (Kling-Gupta efficiency [KGE], Kling et al. 2012). All goodness-of-fit metrics were implemented in the Python and are available from the Pastas software (Collenteur et al. 2019, version 0.17.0).

Table 1
Formulas for the Weighted Goodness-of-Fit Metrics

Metric Name	Equation
MAE (mean absolute error)	$\sum_{i=1}^{N} w_i' \mid x_i - y_i \mid$
RMSE (root mean squared error)	$\sqrt{\sum_{i=1}^{N} w_i'(n_i - \overline{\mu}_n)^2}$
r (Pearson correlation coefficient)	$\frac{\sum_{i=1}^{N} w_i'(x_i - \overline{\mu}_x)(y_i - \overline{\mu}_y)}{\sqrt{\sum_{i=1}^{N} w_i'(x_i - \overline{\mu}_x)^2 \sum_{i=1}^{N} w_i'(y_i - \overline{\mu}_y)^2}}$
R^2 (coefficient of determination)	$1 - \frac{\sum_{i=1}^{N} w_i' n_i^2}{\sum_{i=1}^{N} w_i'(x - \overline{\mu}_x)^2}$
EVP (explained variance percentage)	$\frac{\overline{\sigma}_x^2 - \overline{\sigma}_n^2}{\overline{\sigma}_x^2} \cdot 100\%$
KGE (Kling-Gupta efficiency)	$1 - \sqrt{(r-1)^2 + (\beta - 1)^2 - (\gamma - 1)^2}$ where $\beta = \overline{\mu}_x / \overline{\mu}_y$ and $\gamma = \frac{\overline{\sigma}_x / \overline{\mu}_x}{\overline{\sigma}_y / \overline{\mu}_y}$

Notes: The errors n_i are computed as $n_i = x_i - y_i$. The formula for the weighted MAE is taken from Cleger-Tamayo et al. (2012) and the formula for the weighted r^2 is taken from Bailey et al. (2018).

Benchmarking the Weighted Metrics

The goodness-of-fit metrics were benchmarked using synthetically created "observed" and "simulated" time series, sampled at regular and irregular time intervals. The synthetic time series were constructed as follows:

- In the first step, a time series with regular time steps was constructed consisting of a seasonal trend with a period of 1 year. This time series will be referred to as the simulated time series (e.g., the time series output from a model).
- In the second step, 100 time series with regular time steps were created by adding normally distributed random errors to the simulated time series. Additional errors were added to the first 10 years of the time series to create a change in the variance over the observations period of the time series. These time series are referred to as the regularly observed time series (e.g., the observed time series if measurements were performed at a regular time interval).
- In the third step, each of these 100 time series was sampled using the time indices from 1000 head time series that were observed in the Netherlands between 1993 and 2018. These head time series all consist of a period with low frequency observations and a period with high frequency observations and may contain multiple data gaps (shown in Figure S1). These time series are referred to here as the irregularly observed time series (e.g., the time series observed at irregular time intervals, often found in groundwater data).

This way a total of 100,000 irregular observed time series were created. An example of a synthetic observed time series with regular time steps, irregular time steps, and the simulated time series is shown in Figure 2. For each of the 100,000 time series the goodness-of-fit between the irregularly observed time series and the simulated time series was computed using both the weighted and unweighted metrics. In addition the goodness-of-fit metrics were computed for the fit between the regularly observed and the simulated time series. These are considered to be the reference values of the goodness-of-fit metrics here; the goodness-of-fit between two time series with no missing data. Finally, the ratio between the goodness-of-fit metrics computed using the irregular time series and the reference fit metrics was calculated for all 100,000 time series.

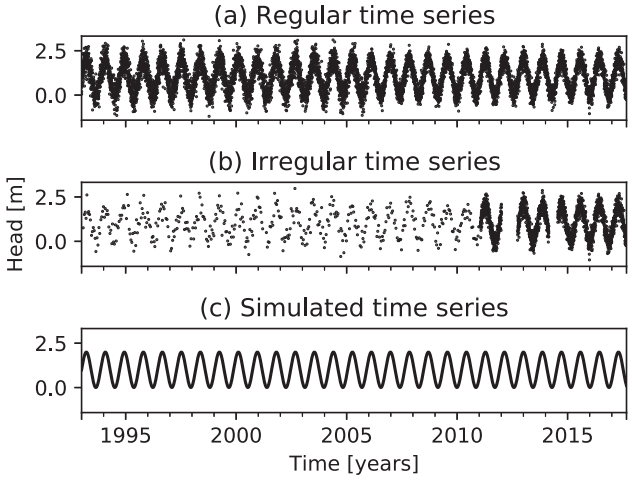

Figure 2. Example of the synthetic time series created for the benchmark test.

The results are summarized in the extended box plots shown in Figure 3. Extended box plots show more quantiles and can also be used to interpret the distribution of the data (Hofmann et al. 2017). A value of one means that the metric computed using the irregular time series is equal to the metric computed using the regular time series. The values of weighted goodness-of-fit metrics are centered around one, indicating that on average the weighted goodness-of-fit metrics equal the reference values. The unweighted metrics all show a systematic bias and generally perform less well than the weighted metrics in this benchmark study. The weighted metrics do show a larger spread around the mean, indicating that the values of these metrics are more uncertain than the unweighted metrics.

In the benchmark test presented above, additional errors were added to the first 10 years of the time series. The unweighted fit metrics thus overestimate the fit between the simulated and the irregularly observed time series because these metrics are biased toward the period with more observations. If the additional errors are added to the final 10 years of the time series, the unweighted metrics would underestimate the fit (shown in Figure S2). In that case, the average fit may thus actually be better than would expected from the computed unweighted fit metric. If no additional errors are added, such that the variance of the errors is constant through

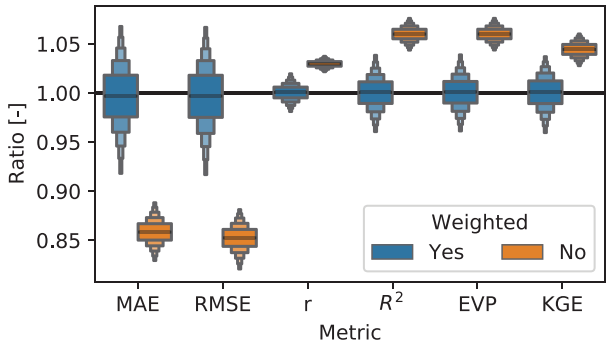

Figure 3. Extended box plots of the ratios between the reference goodness-of-fit metrics and the weighted and unweighted metrics computed for all synthetics time series with irregular time steps.

time, the weighted and unweighted goodness-of-fit metrics will both be unbiased (shown in Figure S3). In this case, the use of unweighted metrics should be preferred because these have smaller uncertainties.

Example Application

To illustrate how the use of weighted goodness-of-fit metrics may impact decisions about the model fit, we return to the hypothetical story from the beginning of this Commentary. Imagine the hydrogeologist created over 100 time series models, but lacks the time to visually inspect the fit of each individual model. She also prefers to use an objective measure for the model fit, and therefore decides to apply the following criterion to decide whether or not a model has a sufficiently good fit to be used for further analysis:

> models with an $R^2 > 0.7$ have a satisfactory model fit.

The hydrogeologist then computes the weighted and unweighted goodness-of-fit metrics. The results for the example data from Figure 1 are shown in Table 2. These clearly show that the model fit is computed to be lower when using weighted metrics (e.g., MAE and RMSE are higher and R^2 and KGE are lower) than when using unweighted metrics. When applying the criterion defined above using the unweighted R^2, the hydrogeologist would conclude that the model can be used for further analysis ($R^2 > 0.7$). The opposite conclusion would be drawn when using the weighted R^2. Similar opposite conclusions would be drawn when applying comparable criteria to the other goodness-of-fit metrics. When visually interpreting the fit of the model shown in Figure 1, one would probably conclude that the model fit is unsatisfactory. This conclusion would be supported by the weighted metrics.

Discussion

The presented goodness-of-fit metrics can be used independent of the approach that is applied to calibrate the model. Whether or not irregular time steps are taken into account during model calibration is, however, likely to influence the difference between weighted and unweighted goodness-of-fit metrics. Taking irregular time steps into account during calibration (see, e.g., Bierkens et al. 1999; Yi and Lee 2004; von Asmuth and Bierkens 2005) should generally help to reduce the differences between weighted and unweighted metrics and is therefore recommended. If irregular time steps are not taken into account during calibration (e.g., all errors are equally weighted), the period with more observations will have a larger weight in the objective function and the model fit is more likely to be biased toward this period. In this case, the unweighted metrics are possibly biased as well and caution is advised when interpreting their values.

The use of weighted goodness-of-fit metrics was proposed here to deal with the irregular time steps commonly found in groundwater time series. A challenge in this approach is to determine the values of the weights that are applied. The time step between observations was used here as a pragmatic solution to this problem. Although this solution showed good results in the benchmark study and a practical application, it would be interesting to see a more formal statistical analysis of the problem in future research. Another approach to explore that would not require the use of weights, is to compute the goodness-of-fit metric over different time periods (e.g., periods of 1 year) and take the average of these. The scripts and data used to benchmark the methods in this study are provided in Appendix S1 to support such future developments.

Table 2
Weighted and Unweighted Goodness-of-Fit Metrics for the Data Shown in Figure 1

	Unweighted	Weighted	Difference
Mean absolute error (m)	0.10	0.23	57.7%
Root mean squared error (m)	0.14	0.29	51.0%
r	0.91	0.64	−41.4%
R^2	0.83	0.34	−139.7%
Explained variance percentage (%)	82.61	39.96	−106.7%
Kling-Gupta efficiency	0.83	0.58	−44.9%

Notes: The metrics are computed over the entire observation period 2000–2018.

Concluding Remarks

In this Technical Commentary it was shown that caution is required when interpreting goodness-of-fit metrics computed for (head) time series with irregular time steps between observations. The use of weighted goodness-of-fit metrics was proposed as a possible solution to obtain goodness-of-fit metrics that are more informative of the average model fit over the entire

observation period. Goodness-of-fit metrics other than those presented can probably be adjusted in a similar fashion to deal with irregular time steps. An example application illustrated how different conclusions about the model fit may be drawn when not taking irregular time steps into account in the computation of such metrics. This is particularly important when numerical values are the primary source of model evaluation; that is, when assessing hundreds of models and using goodness-of-fit metrics as criteria to make a decision if a model is used for further analysis. The use of weighted metrics is recommended as an additional method to detect poorly fitting models in large-scale studies where visual interpretation of individual models is unfeasible.

Acknowledgments

The author acknowledges David Brakenhoff from Artesia for providing the example model that was used to illustrate the problem. The author thanks Christophe Obergfell and one anonymous reviewer for their constructive reviews that helped to improve this manuscript. This work was funded by the Austrian Science Fund (FWF) under Research Grant W1256 (Doctoral Programme Climate Change: Uncertainties, Thresholds, and Coping Strategies).

Author's Note

The author does not have any conflicts of interest or financial disclosures to report.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article. Supporting Information is generally *not* peer reviewed.

Figure S1: Visual presentation of the time stamps at which measurements were obtained for all of the 1000 time series used in the benchmark test.
Figure S2: Results for the benchmark test if additional errors are added to the observed time series for the last 10 years of the time series.
Figure S3: Results for the benchmark test if no additional errors are added to the observed time series.
Appendix S1: Link to Zenodo repository containing all data and scripts used to produce the figures and tables in this manuscript.

References

von Asmuth, J., and M. Bierkens. 2005. Modeling irregularly spaced residual series as a continuous stochastic process. *Water Resources Research* 41, no. 12: W12404. https://doi.org/10.1029/2004WR003726

Bailey, P., A. Emad, T. Zhang, Q. Xie, and E. Sikali. 2018. Weighted and unweighted correlation methods for large-scale educational assessment: wCorr formulas. AIR–NAEP Working Paper No. 2018-01. NCES Data R Project Series #02. ERIC, American Institutes for Research.

Bakker, M., and F. Schaars. 2019. Solving groundwater flow problems with time series analysis: You may not even need another model. *Groundwater* 57, no. 6: 826–833. https://doi.org/10.1111/gwat.12927

Bierkens, M., M. Knotters, and F. van Geer. 1999. Calibration of transfer function–noise models to sparsely or irregularly observed time series. *Water Resources Research* 35, no. 6: 1741–1750. https://doi.org/10.1029/1999WR900083

Cleger-Tamayo, S., J.M. Fernández-Luna, and J.F. Huete. 2012. On the use of weighted mean absolute error in recommender systems. In Proceedings of the Workshop on Recommendation Utility Evaluation *(RUE 2012)*, 24–26. Dublin, Ireland, Citeseer. http://ceur-ws.org/Vol-910/paper5.pdf.

Collenteur, R.A., M. Bakker, R. Caljé, S.A. Klop, and F. Schaars. 2019. Pastas: Open source software for the analysis of groundwater time series. *Groundwater* 57, no. 6: 877–885. https://doi.org/10.1111/gwat.12925

Hofmann, H., H. Wickham, and K. Kafadar. 2017. Letter-value plots: Boxplots for large data. *Journal of Computational and Graphical Statistics* 26, no. 3: 469–477. https://doi.org/10.1080/10618600.2017.1305277

Jackson, E.K., W. Roberts, B. Nelsen, G.P. Williams, E.J. Nelson, and D.P. Ames. 2019. Introductory overview: Error metrics for hydrologic modelling—A review of common practices and an open source library to facilitate use and adoption. *Environmental Modelling & Software* 119: 32–48. https://doi.org/10.1016/j.envsoft.2019.05.001

Kling, H., M. Fuchs, and M. Paulin. 2012. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology* 424–425: 264–277. https://doi.org/10.1016/j.jhydrol.2012.01.011

Yi, M.-J., and K.-K. Lee. 2004. Transfer function-noise modelling of irregularly observed groundwater heads using precipitation data. *Journal of Hydrology* 288, no. 3: 272–287. https://doi.org/10.1016/j.jhydrol.2003.10.020