Comment on "How Good is Your Model Fit? Weighted Goodness-of-Fit Metrics for Irregular Time Series"

Zaadnoordijk, Willem J.

DOI
10.1111/gwat.13175

Publication date
2022

Document Version
Final published version

Published in
Groundwater

Citation (APA)
Zaadnoordijk, W. J. (2022). Comment on "How Good is Your Model Fit? Weighted Goodness-of-Fit Metrics for Irregular Time Series". *Groundwater*, *60*(2), 162-164. https://doi.org/10.1111/gwat.13175

Groundwater

Letter to the Editor/

Comment on "How Good is Your Model Fit? Weighted Goodness-of-Fit Metrics for Irregular Time Series"

Comment by Willem J. Zaadnoordijk[1,2]
[1]TNO Geological Survey of the Netherlands, Princetonlaan 6, Utrecht, The Netherlands
[2]Faculty of Civil Engineering and Geosciences, Water Resources Section, Delft University of Technology, Delft, The Netherlands

Introduction

The technical commentary of Collenteur (2021) touches an important aspect in the use of groundwater head data in the conversion from manual measurements (and sampling) to sensors with automatic dataloggers (e.g., Post and von Asmuth 2013; Retike et al. 2022). Collenteur offers a practical solution that improves evaluation for time series with a transition from regular manual measurements to high(er) frequency automatic logged groundwater heads. The weighting he proposes may also be useful for calibration of time series models. However, scientific underpinning is needed for true advancement in the analysis of such data, and data with other frequency variations. This comment considers the problem from two perspectives: the model Collenteur presented and the head measurements used for the model.

Looking at the Model: Serial Correlation of Residuals

The residuals of the model in Figure 1 of Collenteur clearly have serial correlation, because of the long periods with residues of the same sign so that they are far from randomly distributed. These systematic deviations between model and measurement invalidate the calibration of the model (Hill and Tiedeman 2005), which means that the model may not be used and eliminates the need for evaluation of the differences between model and measurements. Helsel et al. (2020) mention solutions for the problem of serial correlation in the context of linear regression:

1. Sample from the dataset: this assumes the extra measurements in the high frequency part are redundant.
2. Group the data into time periods and compute, for example, a time-weighted mean, and model these means: only applicable with a constant frequency, because the variance of the mean otherwise varies.
3. Add explanatory variables to the model to account for the pattern in time.
4. Use a more sophisticated approach.

If option 1 is used, the extra information provided by the higher frequency is discarded. Option 2 is not applicable because the frequency is not constant, and it would mean modeling with a monthly timestep instead of a daily timestep. Option 3 cannot be used either because there is no potential cause of the deviations. Finally, option 4 includes a logical step: adding a noise model. This should take care of the correlation in the residuals and can do that even when the measurement frequency varies (Bierkens et al. 1999; von Asmuth and Bierkens 2005). However, care must be taken that the implementation of the transfer function noise model does not contain assumptions that are violated in the application to such a series. Examples are simple averaging in the calculation of the constant (e.g., equation 9 in von Asmuth et al. 2002) and in a criterion for innovations (equation 17 in von Asmuth and Bierkens 2005). Instead of simple averaging, weighted averaging should be used taking the considerations of information density and correlation of the measurements into account. The weighting scheme of Collenteur provides a practical solution for this. However, the initial weights are not symmetrical in time:

$$w_i = \min\left(t_i - t_{i-1}, \Delta t_{\max}\right) \quad (1)$$

Received December 2021, accepted January 2022.

doi: 10.1111/gwat.13175

This can be improved by using instead:

$$w_i = \min\left(\frac{(t_i - t_{i-1}) + (t_{i+1} - t_i)}{2}, \Delta t_{\max}\right)$$
$$w_1 = \min\,(t_2 - t_1, \Delta t_{\max})$$
$$w_N = \min\,(t_N - t_{N-1}, \Delta t_{\max}) \qquad (2)$$

These initial weights still need to be normalized and made dimensionless by dividing by the sum of the initial weights (equation 4 in Collenteur 2021) before application.

Collenteur suggests using the timestep of the lowest frequency for $\Delta t_{\max}$. However, a more rigorous approach is needed for usage in model calibration. The response time of the groundwater system or the autocorrelation of the groundwater heads could provide a more physical basis for $\Delta t_{\max}$. This will also make the weighting applicable for series with other frequency variations.

Looking at the Measurements: Information Density and Correlation

Collenteur does not mention correlation—which is obviously present in the high frequency part and to a lesser extent in the low frequency part.

Weights are needed when there is (variable) correlation between measurements (Hill and Tiedeman 2005) to ensure that equal amounts of information have equal weight in a calibration.

The formal solution is a full weight matrix (Hill and Tiedeman 2005). However, this requires information that usually is unknown and thus would require a model. This would lead to an iterative calibration procedure. Also, the matrix can become very large, which makes this approach further impractical.

The effect of correlation is that an individual measurement contains less additional information if the correlation with other measurements is higher. If all measurements are weighed equally, information in the measurements with higher correlation is given more importance than the information from measurements with less correlation. This definitely plays a role in the examples of Collenteur.

However, the solution presented by Collenteur does not account explicitly for correlation, but assumes that the time series contains the same amount of information per period $\Delta t_{\max}$:

$$w_i = \min\,(t_i - t_{i-1}, \Delta t_{\max}) \qquad (3)$$

According to the equation, the information contained in the measurements in the period is independent from the number of measurements and that the higher frequency does not add information per time. However, in the example of Collenteur the information content in the high frequency part does seem higher than in the low frequency part for the selected period $\Delta t_{\max}$ of 1 month, although it is not proportional to the number of measurements in the period due to the higher correlation between two subsequent measurements in the higher frequency part. The correlation between the measurements can only be established objectively with a model. So it cannot be established independently and using the correlation for assigning weights leads to an iterative modeling procedure. In assigning weights, difference in measurement accuracy should be considered also.

Example

As an illustration, I analyzed the time series of the same piezometer (from the Dutch national subsurface information database at https://www.DINOloket.nl/en/) as Collenteur (2021) with precipitation and Makkink evaporation series from the same meteorological stations of the Royal Dutch Meteorological Institute (KNMI) using the Metran software (Berendrecht and van Geer 2016; Zaadnoordijk et al. 2019).

The initial model based on all measurements (orange line in Figure 1) matches the yearly fluctuation reasonably

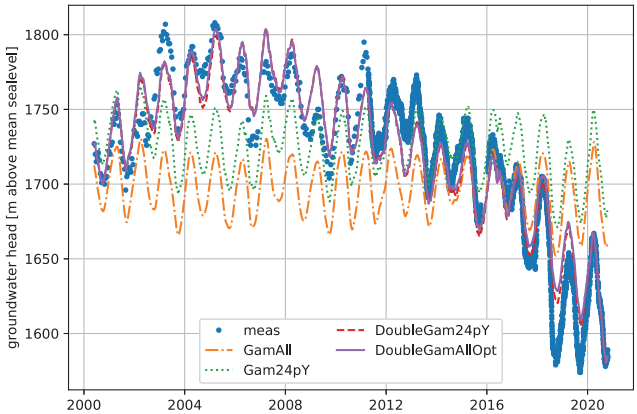

Figure 1. Measurements for piezometer B51F0304012 (blue dots: meas) and various Metran models with daily precipitation (from KNMI precipitation station 908 Deurne) and daily Makkink evaporation (from KNMI meteorological station 375 Volkel), using a single Gamma function on all data (orange dash dot line: GamAll) or on 24 measurements per year (green dotted line: Gam24pY) and using two Gamma functions on 24 measurements per year (red dashed line: DoubleGam24pY) or all measurements (purple line: DoubleGamAllOpt).

well, and the average level reflects more the average of the high frequency part than of the entire series. Next, the frequency of the part with daily measurements has been reduced by selecting only the measurements on the 14th and the 28th day of each month, resulting in 24 measurements per year. The model for this series gives the same fluctuation, but a better average level (green line in Figure 1).

Recognizing that the residuals of the first two models have a multiyear fluctuation that could be due to a much slower response to precipitation and evaporation, a new model has been created in which the responses of the second model have been included with fixed parameters and a second Gamma function has been added for the response of precipitation and evaporation with initial parameters such that the response is slower. This leads to a model that fits the data much better (red line in Figure 1). As a last step, the parameters of the third model have been specified as initial values without fixing any of them and they have been optimized using all measurements (purple line in Figure 1).

This example goes beyond the Commentary of Collenteur on the use of weights in calculating statistics for model evaluation. It shows steps that can be taken to arrive at an acceptable model. Working with a reduced set of measurements, which has a similar effect to the weighting scheme of Collenteur, may help during this model development.

Closing Remark

Weighting proposed by Collenteur is useful in the exploratory phase, but lacks theoretical underpinning and should therefore be avoided for prediction or decision support. Alternative options include the development of structures for the noise model that do a better job of removing autocorrelation in the residuals of a time series model.

ACKNOWLEDGMENT

The example is based on time series modeling carried out in the RESOURCE project within the GeoERA programme, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731166.

References

Berendrecht, W.L., and F.C. van Geer. 2016. A dynamic factor modeling framework for analyzing multiple groundwater head series simultaneously. *Journal of Hydrology* 536: 50–60. http://dx.doi.org/10.1016/j.jhydrol.2016.02.028.

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

Collenteur, R.A. 2021. How good is your model fit? Weighted goodness-of-fit metrics for irregular time series, technical commentary. *Groundwater* 59, no. 4: 474–478. https://doi.org/10.1111/gwat.13111

Helsel, D.R., R.M. Hirsch, R. Ryberg, S.A. Archfield, and E.J. Gilroy. 2020. *Statistical Methods in Water Resources, U.S. Geological Survey Techniques and Methods, Book 4, Chapter A3*. Reston, VA: USGS. https://doi.org/10.3133/tm4a3

Hill, M.C., and C.R. Tiedeman. 2005. *Effective Groundwater Model Calibration: With Analysis of Data, Sensitivities, Predictions, and Uncertainty*. Hoboken, NJ: John Wiley and Sons. https://doi.org/10.1002/0470041080

Post, V.E.A., and J.R. von Asmuth. 2013. Review: Hydraulic head measurements—New technologies, classic pitfalls. *Hydrogeology Journal* 21: 737–750. https://doi.org/10.1007/s10040-013-0969-0

Retike, I., J. Bikše, A. Kalvāns, A. Dēliņa, Z. Avotniece, W.J. Zaadnoordijk, M. Jemeljanova, K. Popovs, A. Babre, A. Zelenkevičs, and A. Baikovs. 2022. Rescue of groundwater level time series: How to visually identify and treat errors. *Journal of Hydrology* 604. https://doi.org/10.1016/j.jhydrol.2021.127294

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

von Asmuth, J.R., M.F.P. Bierkens, and K. Maas. 2002. Transfer function-noise modeling in continuous time using predefined impulse response functions. *Water Resources Research* 38, no. 12. https://doi.org/10.1029/2001WR001136

Zaadnoordijk, W.J., S.A.R. Bus, A. Lourens, and W.L. Berendrecht. 2019. Automated time series modeling for piezometers in the National Database of The Netherlands. *Groundwater* 57, no. 6: 834–843. https://doi.org/10.1111/gwat.12819