Preprint hess-2024-111

Data-driven modeling of hydraulic head time series: results and lessons learned from the 2022 groundwater modeling challenge.

By Raoul A. Collenteur et al. 2024

raoul.collenteur@eawag.ch

This a very interesting paper about modeling groundwater hydraulic head time series.

While the results and conclusions are of significant interest, the modeling performed using BRGM's Gardenia computer code presents a clear concern.

It appears that for the presented project, the users of this computer code have not employed the recommended standard method for modeling hydrological time series. The following issues have been identified:

1. Manual calibration: The code's standard procedure involves automatic calibration. However, in this case, manual calibration was employed without any justification. It is not surprising that this deviation from the standard approach has resulted in inaccurate calibration.
2. Omission of snowmelt module: Even for basins in snow-dominated climates like Sweden, the snowmelt module was not utilized. Consequently, the obtained results are of poor quality.
3. Not using of double-reservoir schemes**:** Double-reservoir schemes are tailored for shallow water level time series, such as the "Netherlands" series. Their absence in this analysis has led to poor simulation of this time series.
4. Disregard of river level integration: The standard feature in the Gardenia computer code for integrating river stage series was not utilized. Using this feature would have significantly improved the results for the "USA" series.

The results presented, resulting from an inappropriate use, strongly discredit BRGM's Gardenia calculation code, which is unacceptable.

We independently modeled the four hydraulic head time series using the data provided in the appendix and achieved satisfactory results:

- In validation phase, the NSE coefficients obtained rank first or second for three out of four wells.
- The average validation NSE rank is 3.25, which is significantly better than the previously presented value of 10.25 (indicating poor performance).

We understand that the paper presents the results from the "2022 groundwater modeling challenge".

However this is our opinion, as having developed Gardenia computer code at BRGM, that this very interesting and valuable paper should be modified to correct the concern of the clear misuse of the model and of the clear discredit on Gardenia model.

Detailed comments:

Line 22: “for the well in the USA, where the lumped-parameter models did not use (or use to the full benefit) the provided river stage data”

Gardenia lumped-parameter model can integrate the provided river stage as an “external influence”. Such an “external influence” is commonly used for the influence of nearby pumping, and also for the variation of river stage or river flow. Taking into account the river stage data for the USA well series significantly improved the NSE criterion during the calibration period: NSE was increased from 0.72 to 0.86

- ⇨ The sentence should be adapted. “most lumped-parameter models, except Gardenia, did not use…”

Line 169: “Gardenia was manually calibrated by minimizing the NSE and visual interpretation.”

This not at all the correct way of using Gardenia. Gardenia, since its creation in 1977, is implemented with an automatic calibration method, the Rosenbrock algorithm. Gardenia is distributed with a tutorial of more than 20 examples, each one with automatic calibration. Gardenia has been used to model more the aquifer level (heads) or the river flow in more than 1000 sites. It has never been calibrated manually.

No wonder than calibrating manually the model leads to poor results.

- ⇨ Our simulations obtained with automatic calibration (computer time between 5 and 10 second for the calibration of each well) will be provided in attached files
- ⇨ The corresponding NSE and MAE criteria will be provided in attached files

Figure 2: Nash-Sutcliffe Efficiency (NSE). The bar plots and ranking of Gardenia do not at all reflect the results obtained with a normal use of the model.

Truly, this discredits this BRGM model (even if it mentioned, line 211 that “none of the models consistently outperformed all other models”)

Indeed after a normal standard automatic calibration of the 4 wells on the calibration period, and then calculating the criteria on the validation period (where the observed heads were totally ignored during the calibration phase), we obtained very different results

Comparing our validation NSE to the NSE values (digitalized) from Figure 2:

Our Gardenia validation phase NSE:

Netherlands	validation NSE = 0.873	=> Rank = **1**, instead of rank 10;
Germany	validation NSE = 0.80	=> Rank = **1** (or 2), instead of rank 8
Sweden	validation NSE = 0.611	=> Rank = **2**, instead of rank 11
USA	validation NSE = 0.862	=> Rank = **9**, instead of rank 12

- ⇨ Average Gardenia rank = **3.25**, instead of rank 10.25 which would be fairly bad.
- ⇨ Gardenia rank = within the two best ranks for 3 wells out 4.
- ⇨ The true bar plot and ranks numbers should be corrected in Figure 2 (and in Figure 4)

Figure 3: Mean Absolute Error (MAE)

Comparing our validation MAE to the MAE values (digitalized) from Figure 2:

Our values of Validation MAE:
Netherlands = 0.057 => Approx rank = **3**, instead of rank 9,
Germany = 0.10 => Approx rank = **4**, instead of rank 10,
Sweden_2 = 0.383 => Approx rank = **2**, instead of rank 11,
USA = 0.255 => => Approx rank = **9**, instead of rank 12

- ⇨ Average Gardenia rank = 4.5, instead of rank 10.5 which would be fairly bad.
- ⇨ The true bar plot and ranks numbers must corrected in this Figure 3.

Line 209: "Model performances generally decreased from the calibration…"

Just for information: our Gardenia modeling: average NSE for the 4 basin:
Calibration 0.807, validation = 0.786 => Very small decrease.

Line 220-224: "Performance of the lumped-parameter models substantially lower for the well in the USA"

In the sentence "The relatively low model performances for HydroSight ~~**and Gardenia**~~ here can probably be explained by the fact that river stage data was not used in these models, opposite to all other teams."

The 2 words "and Gardenia" should be deleted, as using the river stage for the simulation of the USA well, which is standard in Gardenia, yields a very high NSEs: 0. 862 => Rank = 3 for validation, and a very high calibration NSE = 0.893.

Lines 223-226:

"Missing data and processes are likely also the reasons for the low model performance of the Gardenia model for the well in Sweden, i.e., it is the only model in the challenge that did not use temperature data. Temperature data for Sweden is important to account for the impact of snow processes on the heads."

The sentence must be deleted. As a matter of fact, since about 1977 Gardenia is operational with a snow melting module. It make no sense to model a basin (or a well) subject every year to very long periods with negative temperature without using the standard snow melting module. (There are examples of this use in the tutorial provided with the code distribution).

(To our mind, in a lumped parameter model equipped with a snow melting module, disregarding temperature data in such a snow context is as inappropriate as disregarding potential evapotranspiration (PET) data or even precipitation data.)

Using the standard snow melt module, using temperature, for the Sweden_2 well yields satisfying NSEs: 0.611 => Rank = 2 for validation, and 0.777 for calibration.